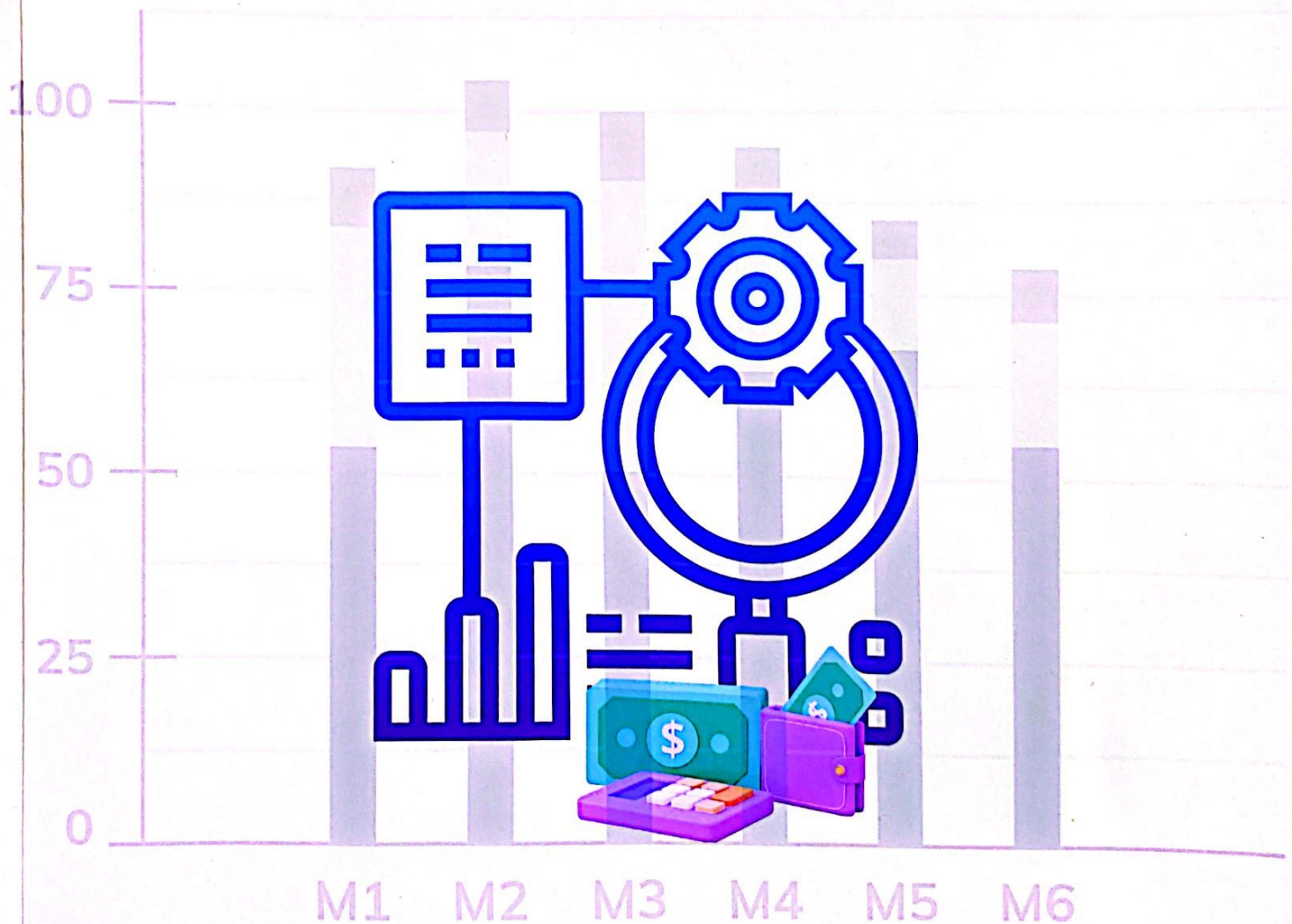


# **Yangon University of Economics**

## **Post Graduate Diploma in Research Studies**

**(9th Batch)**

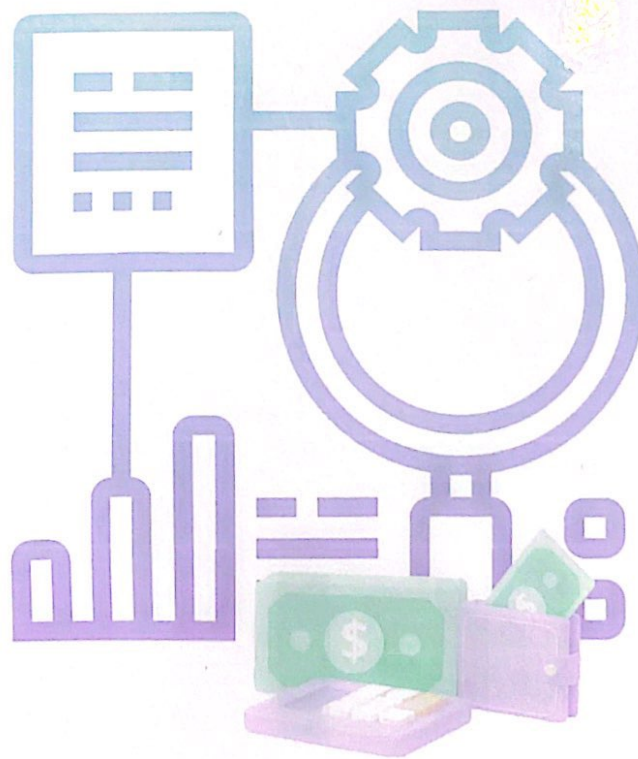


**DRS-112 Exploratory Data Analysis**

**First Quarter**

### 3. Data Description

15. Exploratory data analysis



# Data Description

## STATISTICS TODAY

### How Long Are You Delayed by Road Congestion?

No matter where you live, at one time or another, you have been stuck in traffic. To see whether there are more traffic delays in some cities than in others, statisticians make comparisons using descriptive statistics. A statistical study by the Texas Transportation Institute found that a driver is delayed by road congestion an average of 36 hours per year. To see how selected cities compare to this average, see Statistics Today—Revisited at the end of the chapter.

This chapter will show you how to obtain and interpret descriptive statistics such as measures of average, measures of variation, and measures of position.



Alan Schein/Alamy RF

## OUTLINE

Introduction

**3-1** Measures of Central Tendency

**3-2** Measures of Variation

**3-3** Measures of Position

**3-4** Exploratory Data Analysis

Summary

## OBJECTIVES

After completing this chapter, you should be able to:

- 1** Summarize data, using measures of central tendency, such as the mean, median, mode, and midrange.
- 2** Describe data, using measures of variation, such as the range, variance, and standard deviation.
- 3** Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.
- 4** Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.



## Introduction

Chapter 2 showed how you can gain useful information from raw data by organizing them into a frequency distribution and then presenting the data by using various graphs. This chapter shows the statistical methods that can be used to summarize data. The most familiar of these methods is the finding of averages.

For example, you may read that the average speed of a car crossing midtown Manhattan during the day is 5.3 miles per hour or that the average number of minutes an American father of a 4-year-old spends alone with his child each day is 42.<sup>1</sup>

In the book *American Averages* by Mike Feinsilber and William B. Meed, the authors state:

*“Average” when you stop to think of it is a funny concept. Although it describes all of us it describes none of us. . . . While none of us wants to be the average American, we all want to know about him or her.*

The authors go on to give examples of averages:

*The average American man is five feet, nine inches tall; the average woman is five feet, 3.6 inches.*

*The average American is sick in bed seven days a year missing five days of work.*

*On the average day, 24 million people receive animal bites.*

*By his or her 70th birthday, the average American will have eaten 14 steers, 1050 chickens, 3.5 lambs, and 25.2 hogs.<sup>2</sup>*

In these examples, the word *average* is ambiguous, since several different methods can be used to obtain an average. Loosely stated, the average means the center of the distribution or the most typical case. Measures of average are also called *measures of central tendency* and include the *mean*, *median*, *mode*, and *midrange*.

Knowing the average of a data set is not enough to describe the data set entirely. Even though a shoe store owner knows that the average size of a man’s shoe is size 10, she would not be in business very long if she ordered only size 10 shoes.

### Interesting Fact

A person has an average of 1460 dreams in one year.



© fotog/Getty Images RF

<sup>1</sup>“Harper’s Index,” *Harper’s* magazine.

<sup>2</sup>Mike Feinsilber and William B. Meed, *American Averages* (New York: Bantam Doubleday Dell).

As this example shows, in addition to knowing the average, you must know how the data values are dispersed. That is, do the data values cluster around the average, or are they spread more evenly throughout the distribution? The measures that determine the spread of the data values are called *measures of variation*, or *measures of dispersion*. These measures include the *range*, *variance*, and *standard deviation*.

Finally, another set of measures is necessary to describe data. These measures are called *measures of position*. They tell where a specific data value falls within the data set or its relative position in comparison with other data values. The most common position measures are *percentiles*, *deciles*, and *quartiles*. These measures are used extensively in psychology and education. Sometimes they are referred to as *norms*.

The measures of central tendency, variation, and position explained in this chapter are part of what is called *traditional statistics*.

Section 3–4 shows the techniques of what is called *exploratory data analysis*. These techniques include the *boxplot* and the *five-number summary*. They can be used to explore data to see what they show (as opposed to the traditional techniques, which are used to confirm conjectures about the data).

## 3–1 Measures of Central Tendency

### OBJECTIVE 1

Summarize data, using measures of central tendency, such as the mean, median, mode, and midrange.

Chapter 1 stated that statisticians use samples taken from populations; however, when populations are small, it is not necessary to use samples since the entire population can be used to gain information. For example, suppose an insurance manager wanted to know the average weekly sales of all the company's representatives. If the company employed a large number of salespeople, say, nationwide, he would have to use a sample and make an inference to the entire sales force. But if the company had only a few salespeople, say, only 87 agents, he would be able to use all representatives' sales for a randomly chosen week and thus use the entire population.

Measures found by using all the data values in the population are called *parameters*. Measures obtained by using the data values from samples are called *statistics*; hence, the average of the sales from a sample of representatives is a *statistic*, and the average of sales obtained from the entire population is a *parameter*.

### Historical Note

In 1796, Adolphe Quetelet investigated the characteristics (heights, weights, etc.) of French conscripts to determine the "average man." Florence Nightingale was so influenced by Quetelet's work that she began collecting and analyzing medical records in the military hospitals during the Crimean War. Based on her work, hospitals began keeping accurate records on their patients.

A **statistic** is a characteristic or measure obtained by using the data values from a sample.

A **parameter** is a characteristic or measure obtained by using all the data values from a specific population.

These concepts as well as the symbols used to represent them will be explained in detail in this chapter.

**General Rounding Rule** In statistics the basic rounding rule is that when computations are done in the calculation, rounding should not be done until the final answer is calculated. When rounding is done in the intermediate steps, it tends to increase the difference between that answer and the exact one. But in the textbook and solutions manual, it is not practical to show long decimals in the intermediate calculations; hence, the values in the examples are carried out to enough places (usually three or four) to obtain the same answer that a calculator would give after rounding on the last step.

There are specific rounding rules for many statistics, and they will be given in the appropriate sections.

### The Mean

The *mean*, also known as the *arithmetic average*, is found by adding the values of the data and dividing by the total number of values. For example, the mean of 3, 2, 6, 5, and 4 is found by adding  $3 + 2 + 6 + 5 + 4 = 20$  and dividing by 5; hence, the mean of the data is  $20 \div 5 = 4$ . The values of the data are represented by  $X$ 's. In this data set,  $X_1 = 3$ ,  $X_2 = 2$ ,  $X_3 = 6$ ,

$X_4 = 5$ , and  $X_5 = 4$ . To show a sum of the total  $X$  values, the symbol  $\Sigma$  (the capital Greek letter sigma) is used, and  $\Sigma X$  means to find the sum of the  $X$  values in the data set. The summation notation is explained in the online resource section under “Algebra Review.”

The **mean** is the sum of the values, divided by the total number of values.

The **sample mean**, denoted by  $\bar{X}$  (pronounced “X bar”), is calculated by using sample data. The sample mean is a statistic.

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\Sigma X}{n}$$

where  $n$  represents the total number of values in the sample.

The **population mean**, denoted by  $\mu$  (pronounced “mew”), is calculated by using all the values in the population. The population mean is a parameter.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

where  $N$  represents the total number of values in the population.

In statistics, Greek letters are used to denote parameters, and Roman letters are used to denote statistics. Assume that the data are obtained from samples unless otherwise specified.

**Rounding Rule for the Mean** The mean should be rounded to one more decimal place than occurs in the raw data. For example, if the raw data are given in whole numbers, the mean should be rounded to the nearest tenth. If the data are given in tenths, the mean should be rounded to the nearest hundredth, and so on.

### EXAMPLE 3–1 Avian Flu Cases

The number of confirmed flu cases for a 9-year period is shown. Find the mean.

4   46   98   115   88   44   73   48   62

Source: World Health Organization.

#### SOLUTION

$$\begin{aligned}\bar{X} &= \frac{\Sigma X}{n} = \frac{4 + 46 + 98 + 115 + 88 + 44 + 73 + 48 + 62}{9} \\ &= \frac{578}{9} \approx 64.2\end{aligned}$$

Hence, the mean number of flu cases over the 9-year period is 64.2.

### EXAMPLE 3–2 Store Sales

The data show the systemwide sales (in millions) for U.S. franchises of a well-known donut store for a 5-year period. Find the mean.

\$221   \$239   \$262   \$281   \$318

Source: Krispy Kreme.

#### SOLUTION

$$\bar{X} = \frac{\Sigma X}{n} = \frac{221 + 239 + 262 + 281 + 318}{5} = \frac{1321}{5} = 264.2$$

The mean amount of sales for the stores over the 5-year period is \$264.2 million.

The mean, in most cases, is not an actual data value.

The procedure for finding the mean for grouped data assumes that the mean of all the raw data values in each class is equal to the midpoint of the class. In reality, this is not true,

since the average of the raw data values in each class usually will not be exactly equal to the midpoint. However, using this procedure will give an acceptable approximation of the mean, since some values fall above the midpoint and other values fall below the midpoint for each class, and the midpoint represents an estimate of all values in the class.

The steps for finding the mean for grouped data are shown in the next Procedure Table.

### Procedure Table

#### Finding the Mean for Grouped Data

**Step 1** Make a table as shown.

A Class	B Frequency $f$	C Midpoint $X_m$	D $f \cdot X_m$
------------	--------------------	---------------------	--------------------

**Step 2** Find the midpoints of each class and place them in column C.

**Step 3** Multiply the frequency by the midpoint for each class, and place the product in column D.

**Step 4** Find the sum of column D.

**Step 5** Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n}$$

[Note: The symbols  $\Sigma f \cdot X_m$  mean to find the sum of the product of the frequency ( $f$ ) and the midpoint ( $X_m$ ) for each class.]

### EXAMPLE 3–3 Salaries of CEOs

The frequency distribution shows the salaries (in millions) for a specific year of the top 25 CEOs in the United States. Find the mean.

Source: S & P Capital.

Class boundaries	Frequency
15.5–20.5	13
20.5–25.5	6
25.5–30.5	4
30.5–35.5	1
35.5–40.5	1
Total	25

#### SOLUTION

A Class	B Frequency	C Midpoint $X_m$	D $f \cdot X_m$
15.5–20.5	13		
20.5–25.5	6		
25.5–30.5	4		
30.5–35.5	1		
35.5–40.5	1		
	$n = 25$		

#### Interesting Fact

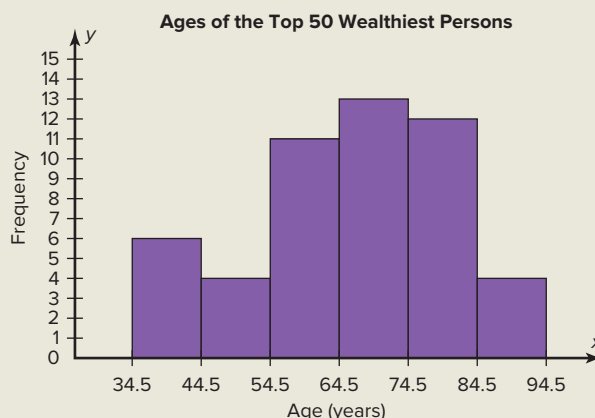
The average time it takes a person to find a new job is 5.9 months.

The histogram shows the ages of the top 50 wealthiest individuals according to *Forbes Magazine* for a recent year. The mean age is 66 years. The median age is

68 years. Explain why these two statistics are not enough to adequately describe the data.



© Don Farrall/Getty Images RF



### Unusual Stat

A person looks, on average, at about 14 homes before he or she buys one.

$$X_m = \frac{15.5 + 20.5}{2} = 18 \quad \frac{20.5 + 25.5}{2} = 23 \quad \text{etc.}$$

$$13 \cdot 18 = 234 \quad 6 \cdot 23 = 138 \quad \text{etc.}$$

A Class	B Frequency	C Midpoint $X_m$	D $f \cdot X_m$
15.5–20.5	13	18	234
20.5–25.5	6	23	138
25.5–30.5	4	28	112
30.5–35.5	1	33	33
35.5–40.5	1	38	38
	$n = 25$		$\Sigma f \cdot X_m = 555$

$$\bar{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{555}{25} = \$22.2 \text{ million}$$

The mean salary is \$22.2 million.

### Historical Note

The concept of median was used by Gauss at the beginning of the 19th century and introduced as a statistical concept by Francis Galton around 1874. The mode was first used by Karl Pearson in 1894.

### The Median

An article recently reported that the median income for college professors was \$43,250. This measure of central tendency means that one-half of all the professors surveyed earned more than \$43,250, and one-half earned less than \$43,250.

The *median* is the halfway point in a data set. Before you can find this point, the data must be arranged in ascending or increasing order. When the data set is ordered, it is called a **data array**. The median either will be a specific value in the data set or will fall between two values, as shown in the next examples.





Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

The median number of tornadoes is 810.

## The Mode

The third measure of average is called the *mode*. The mode is the value that occurs most often in the data set. It is sometimes said to be the most typical case.

The value that occurs most often in a data set is called the **mode**.

A data set that has only one value that occurs with the greatest frequency is said to be **unimodal**.

If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal**. If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal**. When no data value occurs more than once, the data set is said to have *no mode*. *Note: Do not say that the mode is zero.* That would be incorrect, because in some data, such as temperature, zero can be an actual value. A data set can have more than one mode or no mode at all. These situations will be shown in some of the examples that follow.

### EXAMPLE 3-6 Public Libraries

The data show the number of public libraries in a sample of eight states. Find the mode.

114 77 21 101 311 77 159 382

Source: *The World Almanac*.

#### SOLUTION

It is helpful to arrange the data in order, although it is not necessary.

21 77 77 101 114 159 311 382

Since 77 occurs twice, a frequency larger than that of any other number, the mode is 77.

### EXAMPLE 3-7 Licensed Nuclear Reactors

The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

Source: *The World Almanac and Book of Facts*.

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

#### SOLUTION

Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

**EXAMPLE 3–8 U.S. Patent Leaders**

The data show the number of patents secured for the top 5 companies for a specific year. Find the mode.

6180 4894 2821 2559 2483

Source: IFI Claims Patent Services.

Since each value occurs only once, there is no mode.

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

**EXAMPLE 3–9 Salaries of CEOs**

Find the modal class for the frequency distribution for the salaries of the top CEOs in the United States, shown in Example 3–3.

**SOLUTION**

Class	Frequency
15.5–20.5	13 ← Modal class
20.5–25.5	6
25.5–30.5	4
30.5–35.5	1
35.5–40.5	1

Since the class 15.5–20.5 has the largest frequency, 13, it is the modal class. Sometimes the midpoint of the class is used. In this case, it is 18.

The mode is the only measure of central tendency that can be used in finding the most typical case when the data are nominal or categorical.

**EXAMPLE 3–10 Nonalcoholic Beverages**

The data show the number of gallons of various nonalcoholic drinks Americans consume in a year. Find the mode.

Drink	Gallons
Soft drinks	52
Water	34
Milk	26
Coffee	21

Source: U.S. Department of Agriculture.

**SOLUTION**

Since the category of soft drinks has the largest frequency, 52, we can say that the mode or most typical drink is a soft drink.

An extremely high or extremely low data value in a data set can have a striking effect on the mean of the data set. These extreme values are called *outliers*. This is one reason why, when analyzing a frequency distribution, you should be aware of any of these values. For the data set shown in Example 3–11, the mean, median, and mode can be quite different because of extreme values. A method for identifying outliers is given in Section 3–3.

### EXAMPLE 3–11 Salaries of Personnel

A small company consists of the owner, the manager, the salesperson, and two technicians, all of whose annual salaries are listed here. (Assume that this is the entire population.)

Staff	Salary
Owner	\$100,000
Manager	40,000
Salesperson	24,000
Technician	18,000
Technician	18,000

Find the mean, median, and mode.

#### SOLUTION

$$\mu = \frac{\sum X}{N} = \frac{\$100,000 + 40,000 + 24,000 + 18,000 + 18,000}{5} = \frac{\$200,000}{5} = \$40,000$$

Hence, the mean is \$40,000, the median is \$24,000, and the mode is \$18,000.

In Example 3–11, the mean is much higher than the median or the mode. This is so because the extremely high salary of the owner tends to raise the value of the mean. In this and similar situations, the median should be used as the measure of central tendency.

### The Midrange

The *midrange* is a rough estimate of the middle. It is found by adding the lowest and highest values in the data set and dividing by 2. It is a very rough estimate of the average and can be affected by one extremely high or low value.

The **midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$MR = \frac{\text{lowest value} + \text{highest value}}{2}$$

### EXAMPLE 3–12 Bank Failures

The number of bank failures for a recent five-year period is shown. Find the midrange.

3, 30, 148, 157, 71

Source: Federal Deposit Insurance Corporation.

#### SOLUTION

The lowest data value is 3, and the highest data value is 157.

$$MR = \frac{3 + 157}{2} = \frac{160}{2} = 80$$

The midrange for the number of bank failures is 80.



**EXAMPLE 3-13** NFL Signing Bonuses

Find the midrange of data for the NFL signing bonuses. The bonuses in millions of dollars are

18, 14, 34.5, 10, 11.3, 10, 12.4, 10

**SOLUTION**

The lowest bonus is \$10 million, and the largest bonus is \$34.5 million.

$$\text{MR} = \frac{10 + 34.5}{2} = \frac{44.5}{2} = \$22.25 \text{ million}$$

Notice that this amount is larger than seven of the eight amounts and is not typical of the average of the bonuses. The reason is that there is one very high bonus, namely, \$34.5 million.

In statistics, several measures can be used for an average. The most common measures are the mean, median, mode, and midrange. Each has its own specific purpose and use. Exercises 36 through 38 show examples of other averages, such as the harmonic mean, the geometric mean, and the quadratic mean. Their applications are limited to specific areas, as shown in the exercises.

**The Weighted Mean**

Sometimes, you must find the mean of a data set in which not all values are equally represented. Consider the case of finding the average cost of a gallon of gasoline for three taxis. Suppose the drivers buy gasoline at three different service stations at a cost of \$3.22, \$3.53, and \$3.63 per gallon. You might try to find the average by using the formula

$$\begin{aligned}\bar{X} &= \frac{\sum X}{n} \\ &= \frac{3.22 + 3.53 + 3.63}{3} = \frac{10.38}{3} = \$3.46\end{aligned}$$

But not all drivers purchased the same number of gallons. Hence, to find the true average cost per gallon, you must take into consideration the number of gallons each driver purchased.

The type of mean that considers an additional factor is called the *weighted mean*, and it is used when the values are not all equally represented.

Find the **weighted mean** of a variable  $X$  by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\bar{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum wX}{\sum w}$$

where  $w_1, w_2, \dots, w_n$  are the weights and  $X_1, X_2, \dots, X_n$  are the values.

Example 3-14 shows how the weighted mean is used to compute a grade point average. Since courses vary in their credit value, the number of credits must be used as weights.

**EXAMPLE 3-14** Grade Point Average

A student received an A in English Composition I (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology I (4 credits), and a D in Physical Education (2 credits). Assuming A = 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

**SOLUTION****Unusual Stat**

Of people in the United States, 45% live within 15 minutes of their best friend.

Course	Credits ( $w$ )	Grade ( $X$ )
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\bar{X} = \frac{\sum wX}{\sum w} = \frac{3 \cdot 4 + 3 \cdot 2 + 4 \cdot 3 + 2 \cdot 1}{3 + 3 + 4 + 2} = \frac{32}{12} \approx 2.7$$

The grade point average is 2.7.

**TABLE 3–1 Summary of Measures of Central Tendency**

Measure	Definition	Symbol(s)
Mean	Sum of values, divided by total number of values	$\mu, \bar{X}$
Median	Middle point in data set that has been ordered	MD
Mode	Most frequent data value	None
Midrange	Lowest value plus highest value, divided by 2	MR

Table 3–1 summarizes the measures of central tendency.

Researchers and statisticians must know which measure of central tendency is being used and when to use each measure of central tendency. The properties and uses of the four measures of central tendency are summarized next.

**Properties and Uses of Central Tendency****The Mean**

1. The mean is found by using all the values of the data.
2. The mean varies less than the median or mode when samples are taken from the same population and all three measures are computed for these samples.
3. The mean is used in computing other statistics, such as the variance.
4. The mean for the data set is unique and not necessarily one of the data values.
5. The mean cannot be computed for the data in a frequency distribution that has an open-ended class.
6. The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.

**The Median**

1. The median is used to find the center or middle value of a data set.
2. The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
3. The median is used for an open-ended distribution.
4. The median is affected less than the mean by extremely high or extremely low values.

**The Mode**

1. The mode is used when the most typical case is desired.
2. The mode is the easiest average to compute.
3. The mode can be used when the data are nominal or categorical, such as religious preference, gender, or political affiliation.

- The mode is not always unique. A data set can have more than one mode, or the mode may not exist for a data set.

#### The Midrange

- The midrange is easy to compute.
- The midrange gives the midpoint.
- The midrange is affected by extremely high or low values in a data set.

### Distribution Shapes

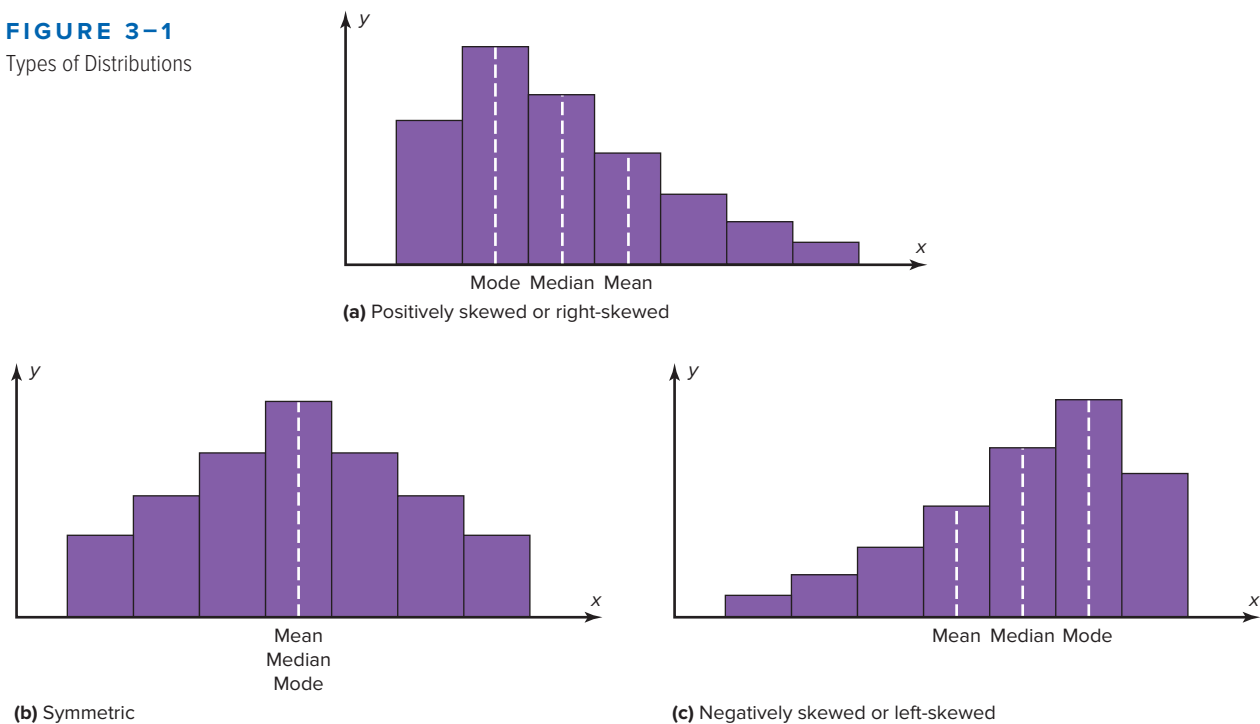
Frequency distributions can assume many shapes. The three most important shapes are positively skewed, symmetric, and negatively skewed. Figure 3-1 shows histograms of each.

In a **positively skewed** or **right-skewed distribution**, the majority of the data values fall to the left of the mean and cluster at the lower end of the distribution; the “tail” is to the right. Also, the mean is to the right of the median, and the mode is to the left of the median.

For example, if an instructor gave an examination and most of the students did poorly, their scores would tend to cluster on the left side of the distribution. A few high scores would constitute the tail of the distribution, which would be on the right side. Another example of a positively skewed distribution is the incomes of the population of the United States. Most of the incomes cluster about the low end of the distribution; those with high incomes are in the minority and are in the tail at the right of the distribution.

In a **symmetric distribution**, the data values are evenly distributed on both sides of the mean. In addition, when the distribution is unimodal, the mean, median, and mode are the same and are at the center of the distribution. Examples of symmetric distributions are IQ scores and heights of adult males.

**FIGURE 3-1**  
Types of Distributions



When the majority of the data values fall to the right of the mean and cluster at the upper end of the distribution, with the tail to the left, the distribution is said to be **negatively skewed** or **left-skewed**. Also, the mean is to the left of the median, and the mode is to the right of the median. As an example, a negatively skewed distribution results if the majority of students score very high on an instructor's examination. These scores will tend to cluster to the right of the distribution.

When a distribution is extremely skewed, the value of the mean will be pulled toward the tail, but the majority of the data values will be greater than the mean or less than the mean (depending on which way the data are skewed); hence, the median rather than the mean is a more appropriate measure of central tendency. An extremely skewed distribution can also affect other statistics.

A measure of skewness for a distribution is discussed in Exercise 48 in Section 3–2.

### Applying the Concepts 3–1

#### Teacher Salaries

The following data from several years ago represent salaries (in dollars) from a school district in Greenwood, South Carolina.

10,000	11,000	11,000	12,500	14,300	17,500
18,000	16,600	19,200	21,560	16,400	107,000

1. First, assume you work for the school board in Greenwood and do not wish to raise taxes to increase salaries. Compute the mean, median, and mode, and decide which one would best support your position to not raise salaries.
2. Second, assume you work for the teachers' union and want a raise for the teachers. Use the best measure of central tendency to support your position.
3. Explain how outliers can be used to support one or the other position.
4. If the salaries represented every teacher in the school district, would the averages be parameters or statistics?
5. Which measure of central tendency can be misleading when a data set contains outliers?
6. When you are comparing the measures of central tendency, does the distribution display any skewness? Explain.

See page 184 for the answers.

### Exercises 3–1

- 1. Roller Coaster Speeds** The data show the heights in feet of 14 roller coasters. Find the mean, median, midrange, and mode for the data.

95	105	50	125	102	120	160
102	118	91	160	95	50	84

Source: UltimateRollerCoaster.com.

- 2. Airport Parking** The number of short-term parking spaces at 15 airports is shown. Find the mean, median, mode, and midrange for the data.

750	3400	1962	700	203
900	8662	260	1479	5905
9239	690	9822	1131	2516

Source: USA Today.

- 3. Length of School Years** The lengths of school years in a sample of various countries in the world are shown. Find the mean, median, midrange, and mode of the data.

251	243	226	216	196	180
-----	-----	-----	-----	-----	-----

Source: U.S. News and World Report.



- 4. Observers in the Frogwatch Program** The number of observers in the Frogwatch USA program (a wildlife conservation program dedicated to helping conserve frogs and toads) for the top 10 states with the most observers is 484, 483, 422, 396, 378, 352, 338, 331, 318, and 302. The top 10 states with the most active watchers list these numbers of visits: 634, 464, 406, 267, 219, 194, 191, 150, 130, and 114. Find the mean, median, mode, and midrange for the data. Compare the measures of central tendency for these two groups of data.

Source: [www.nwf.org/frogwatch](http://www.nwf.org/frogwatch)

- 5. Top Video Games** The following represent XBOX One Top Selling Games and units sold:

Titanfall	2,000,000
Call of Duty: Ghosts	1,790,000
Battlefield 4	1,340,000
Forza Motorsport 5	1,340,000
Tomb Rider: Definitive Edition	1,210,000
Dead Rising 3	1,100,000
Metal Gear Solid 5: Ground Zeroes	980,000
Assassins Creed 4: Black Flag	310,300
Madden NFL 25	303,000
Metro Redux	298,000

Find the mean, median, mode, and midrange.

Source: Statistic Brain Research Institute

- 6. Earnings of Nonliving Celebrities** *Forbes* magazine prints an annual Top-Earning Nonliving Celebrities list (based on royalties and estate earnings). Find the mean, median, mode, and midrange for the data. Comment on the skewness. Figures represent millions of dollars.

Kurt Cobain	50	Ray Charles	10
Elvis Presley	42	Marilyn Monroe	8
Charles M. Schulz	35	Johnny Cash	8
John Lennon	24	J.R.R. Tolkien	7
Albert Einstein	20	George Harrison	7
Andy Warhol	19	Bob Marley	7
Theodore Geisel	10		
(Dr. Seuss)			

Source: [articles.moneycentral.msn.com](http://articles.moneycentral.msn.com)

- 7. Paid Days Off** The data show the number of paid days off workers get in a sample of various countries of the world. Find the mean, median, midrange, and mode for the data.

38	29	10	34	28	30
30	26	19	20	25	

Source: Center for Economic and Policy Research.

- 8. Top-Paid CEOs** The data shown are the total compensation (in millions of dollars) for the 50 top-paid CEOs for a recent year. Compare the averages, and state which one you think is the best measure.

17.5	18.0	36.8	31.7	31.7
17.3	24.3	47.7	38.5	17.0
23.7	16.5	25.1	17.4	18.0
37.6	19.7	21.4	28.6	21.6
19.3	20.0	16.9	25.2	19.8
25.0	17.2	20.4	20.1	29.1
19.1	25.2	23.2	25.9	24.0
41.7	24.0	16.8	26.8	31.4
16.9	17.2	24.1	35.2	19.1
22.9	18.2	25.4	35.4	25.5

Source: *USA TODAY*.

- 9. Airline Passengers** The data show a sample of the number of passengers in millions that major airlines carried for a recent year. Find the mean, median, midrange, and mode for the data.

143.8	17.7	8.5	120.4	33.0	7.1	27.1
10.0	5.0	6.1	4.3	3.1	12.1	

Source: Airlines for America.

- 10. Foreign Workers** The number of foreign workers' certificates for the New England states and the northwestern states is shown. Find the mean, median, and mode for both areas and compare the results.

New England states	Northwestern states
6768	1870
3196	622
1112	620
819	23
1019	172
1795	112

Source: Department of Labor.

- 11. Distances of Stars** Of the 25 brightest stars, the distances from earth (in light-years) for those with distances less than 100 light-years are found below. Find the mean, median, mode, and midrange for the data.

8.6	36.7	42.2	16.8	33.7	77.5	87.9
4.4	25.3	11.4	65.1	25.1	51.5	

Source: *New York Times Almanac 2010*.

- 12. Contest Spelling Words** The last words given and spelled correctly at the National Spelling Bee for the past 21 years are spelled out below. Count the number of letters in each word, and find the mean, median, mode, and midrange for the data.

fibranne	euonym	autochthonous
antipyretic	chiaroscurist	appoggiatura
lyceum	logorrhea	Ursprache
kamikaze	demarche	serrefine
antediluvian	succedaneum	guerdon
xanthosis	propispicience	stromuhr
vivisepulture	pococurante	cymatrichous

Source: *World Almanac 2012*.

- 13. Wind Speeds** The data show the maximum wind speeds for a sample of 40 states. Find the mean and modal class for the data.

Class boundaries	Frequency
47.5–54.5	3
54.5–61.5	2
61.5–68.5	9
68.5–75.5	13
75.5–82.5	8
82.5–89.5	3
89.5–96.5	2

Source: NOAA

- 14. Hourly Compensation for Production Workers** The hourly compensation costs (in U.S. dollars) for production workers in selected countries are represented below. Find the mean and modal class for the data.

Class	Frequency
2.48–7.48	7
7.49–12.49	3
12.50–17.50	1
17.51–22.51	7
22.52–27.52	5
27.53–32.53	5

Compare the mean of these grouped data to the U.S. mean of \$21.97.

Source: New York Times Almanac.

- 15. Points in Rose Bowl Games** The data show the number of points the winning team scored in the Rose Bowl. Find the mean and modal class for the data.

Class	Frequency
14–20	10
21–27	11
28–34	6
35–41	8
42–48	4
49–55	1

Source: The World Almanac and Book of Facts.

- 16. Percentage of Foreign-Born People** The percentage of foreign-born population for each of the 50 states is represented here. Find the mean and modal class for the data. Do you think the mean is the best average for this set of data? Explain.

Percentage	Frequency
0.8–4.4	26
4.5–8.1	11
8.2–11.8	4
11.9–15.5	5
15.6–19.2	2
19.3–22.9	1
23.0–26.6	1

Source: World Almanac.

- 17. Percentage of College-Educated Population over 25** Below are the percentages of the population over 25 years of age who have completed 4 years of college or more for the 50 states and the District of Columbia. Find the mean and modal class.

Percentage	Frequency
15.2–19.6	3
19.7–24.1	15
24.2–28.6	19
28.7–33.1	6
33.2–37.6	7
37.7–42.1	0
42.2–46.6	1

Source: New York Times Almanac.

- 18. Net Worth of Corporations** These data represent the net worth (in millions of dollars) of 45 national corporations. Find the mean and modal class for the data.

Class limits	Frequency
10–20	2
21–31	8
32–42	15
43–53	7
54–64	10
65–75	3

- 19. Specialty Coffee Shops** A random sample of 30 states shows the number of specialty coffee shops for a specific company. Find the mean and modal class for the data.

Class boundaries	Frequency
0.5–19.5	12
19.5–38.5	7
38.5–57.5	5
57.5–76.5	3
76.5–95.5	3

- 20. Commissions Earned** This frequency distribution represents the commission earned (in dollars) by 100 salespeople employed at several branches of a large chain store. Find the mean and modal class for the data.

Class limits	Frequency
150–158	5
159–167	16
168–176	20
177–185	21
186–194	20
195–203	15
204–212	3

- 21. Children of U.S. Presidents** The data show the number of children U.S. Presidents through Obama, had. Construct an ungrouped frequency distribution and find the mean and modal class.

0	5	6	0	3	4	1	4	10	8
7	0	6	1	0	3	4	5	4	8
7	3	5	2	1	2	1	5	3	3
0	0	2	2	6	1	2	3	2	2
4	4	2	6	1	2	2			

Source: World Almanac and Book of Facts.

- 22. Enrollments for Selected Independent Religiously Controlled 4-Year Colleges** Listed below are the enrollments for selected independent religiously controlled 4-year colleges that offer bachelor's degrees only. Construct a grouped frequency distribution with six classes and find the mean and modal class.

1013 1867 1268 1666 2309 1231 3005 2895 2166 1136  
1532 1461 1750 1069 1723 1827 1155 1714 2391 2155  
1412 1688 2471 1759 3008 2511 2577 1082 1067 1062  
1319 1037 2400

Source: World Almanac.

- 23. Automobile Selling Prices** Find the weighted mean price of three models of automobiles sold. The number and price of each model sold are shown in this list.

Model	Number	Price
A	8	\$10,000
B	10	12,000
C	12	8,000

- 24. Fat Grams** Using the weighted mean, find the average number of grams of fat per ounce of meat or fish that a person would consume over a 5-day period if he ate these:

Meat or fish	Fat (g/oz)
3 oz fried shrimp	3.33
3 oz veal cutlet (broiled)	3.00
2 oz roast beef (lean)	2.50
2.5 oz fried chicken drumstick	4.40
4 oz tuna (canned in oil)	1.75

Source: The World Almanac and Book of Facts.

- 25. Diet Cola Preference** A recent survey of a new diet cola reported the following percentages of people who liked the taste. Find the weighted mean of the percentages.

Area	% Favored	Number surveyed
1	40	1000
2	30	3000
3	50	800

- 26. Costs of Helicopters** The costs of three models of helicopters are shown here. Find the weighted mean of the costs of the models.

Model	Number sold	Cost
Sunscaper	9	\$427,000
Skycoaster	6	365,000
High-flyer	12	725,000

- 27. Final Grade** An instructor grades exams, 20%; term paper, 30%; final exam, 50%. A student had grades of 83, 72, and 90, respectively, for exams, term paper, and final exam. Find the student's final average. Use the weighted mean.

- 28. Final Grade** Another instructor gives four 1-hour exams and one final exam, which counts as two 1-hour exams. Find a student's grade if she received 62, 83, 97, and 90 on the 1-hour exams and 82 on the final exam.

- 29.** For these situations, state which measure of central tendency—mean, median, or mode—should be used.

- The most typical case is desired.
- The distribution is open-ended.
- There is an extreme value in the data set.
- The data are categorical.
- Further statistical computations will be needed.
- The values are to be divided into two approximately equal groups, one group containing the larger values and one containing the smaller values.

- 30.** Describe which measure of central tendency—mean, median, or mode—was probably used in each situation.

- One-half of the factory workers make more than \$5.37 per hour, and one-half make less than \$5.37 per hour.
- The average number of children per family in the Plaza Heights Complex is 1.8.
- Most people prefer red convertibles over any other color.
- The average person cuts the lawn once a week.
- The most common fear today is fear of speaking in public.
- The average age of college professors is 42.3 years.

- 31.** What types of symbols are used to represent sample statistics? Give an example. What types of symbols are used to represent population parameters? Give an example.

- 32.** A local fast-food company claims that the average salary of its employees is \$13.23 per hour. An employee states that most employees make minimum wage. If both are being truthful, how could both be correct?

## Extending the Concepts

- 33.** If the mean of five values is 64, find the sum of the values.
- 34.** If the mean of five values is 8.2 and four of the values are 6, 10, 7, and 12, find the fifth value.

- 35.** Find the mean of 10, 20, 30, 40, and 50.

- Add 10 to each value and find the mean.
- Subtract 10 from each value and find the mean.
- Multiply each value by 10 and find the mean.

- d. Divide each value by 10 and find the mean.
- e. Make a general statement about each situation.

**36. Harmonic Mean** The *harmonic mean* (HM) is defined as the number of values divided by the sum of the reciprocals of each value. The formula is

$$HM = \frac{n}{\sum(1/X)}$$

For example, the harmonic mean of 1, 4, 5, and 2 is

$$HM = \frac{4}{1/1 + 1/4 + 1/5 + 1/2} \approx 2.051$$

This mean is useful for finding the average speed. Suppose a person drove 100 miles at 40 miles per hour and returned driving 50 miles per hour. The average miles per hour is *not* 45 miles per hour, which is found by adding 40 and 50 and dividing by 2. The average is found as shown.

Since

$$\text{Time} = \text{distance} \div \text{rate}$$

then

$$\text{Time 1} = \frac{100}{40} = 2.5 \text{ hours to make the trip}$$

$$\text{Time 2} = \frac{100}{50} = 2 \text{ hours to return}$$

Hence, the total time is 4.5 hours, and the total miles driven are 200. Now, the average speed is

$$\text{Rate} = \frac{\text{distance}}{\text{time}} = \frac{200}{4.5} \approx 44.444 \text{ miles per hour}$$

This value can also be found by using the harmonic mean formula

$$HM = \frac{2}{1/40 + 1/50} \approx 44.444$$

Using the harmonic mean, find each of these.

- a. A salesperson drives 300 miles round trip at 30 miles per hour going to Chicago and 45 miles per hour returning home. Find the average miles per hour.
- b. A bus driver drives the 50 miles to West Chester at 40 miles per hour and returns driving 25 miles per hour. Find the average miles per hour.
- c. A carpenter buys \$500 worth of nails at \$50 per pound and \$500 worth of nails at \$10 per pound. Find the average cost of 1 pound of nails.

**37. Geometric Mean** The *geometric mean* (GM) is defined as the  $n$ th root of the product of  $n$  values. The formula is

$$GM = \sqrt[n]{(X_1)(X_2)(X_3) \cdots (X_n)}$$

The geometric mean of 4 and 16 is

$$GM = \sqrt{(4)(16)} = \sqrt{64} = 8$$

The geometric mean of 1, 3, and 9 is

$$GM = \sqrt[3]{(1)(3)(9)} = \sqrt[3]{27} = 3$$

The geometric mean is useful in finding the average of percentages, ratios, indexes, or growth rates. For example, if a person receives a 20% raise after 1 year of service and a 10% raise after the second year of service, the average percentage raise per year is not 15 but 14.89%, as shown.

$$GM = \sqrt{(1.2)(1.1)} \approx 1.1489$$

or

$$GM = \sqrt{(120)(110)} \approx 114.89\%$$

His salary is 120% at the end of the first year and 110% at the end of the second year. This is equivalent to an average of 14.89%, since  $114.89\% - 100\% = 14.89\%$ .

This answer can also be shown by assuming that the person makes \$10,000 to start and receives two raises of 20% and 10%.

$$\text{Raise 1} = 10,000 \cdot 20\% = \$2000$$

$$\text{Raise 2} = 12,000 \cdot 10\% = \$1200$$

His total salary raise is \$3200. This total is equivalent to

$$\$10,000 \cdot 14.89\% = \$1489.00$$

$$\$11,489 \cdot 14.89\% = \$1710.71$$

$$\$3199.71 \approx \$3200$$

Find the geometric mean of each of these.

- a. The growth rates of the Living Life Insurance Corporation for the past 3 years were 35, 24, and 18%.
- b. A person received these percentage raises in salary over a 4-year period: 8, 6, 4, and 5%.
- c. A stock increased each year for 5 years at these percentages: 10, 8, 12, 9, and 3%.
- d. The price increases, in percentages, for the cost of food in a specific geographic region for the past 3 years were 1, 3, and 5.5%.

**38. Quadratic Mean** A useful mean in the physical sciences (such as voltage) is the *quadratic mean* (QM), which is found by taking the square root of the average of the squares of each value. The formula is

$$QM = \sqrt{\frac{\sum X^2}{n}}$$

The quadratic mean of 3, 5, 6, and 10 is

$$\begin{aligned} QM &= \sqrt{\frac{3^2 + 5^2 + 6^2 + 10^2}{4}} \\ &= \sqrt{42.5} \approx 6.519 \end{aligned}$$

Find the quadratic mean of 8, 6, 3, 5, and 4.

**39. Median for Grouped Data** An approximate median can be found for data that have been grouped into a frequency distribution. First it is necessary to find the median class. This is the class that contains the median value. That is the  $n/2$  data value. Then it is assumed that



the data values are evenly distributed throughout the median class. The formula is

$$MD = \frac{n/2 - cf}{f}(w) + L_m$$

where  $n$  = sum of frequencies  
 $cf$  = cumulative frequency of class immediately preceding the median class

$w$  = width of median class

$f$  = frequency of median class

$L_m$  = lower boundary of median class

Using this formula, find the median for data in the frequency distribution of Exercise 16.

## Technology

### EXCEL

#### Step by Step

## Step by Step

### Finding Measures of Central Tendency

#### Example XL3–1

Find the mean, mode, and median of the data from Example 3–7. The data represent the population of licensed nuclear reactors in the United States for a recent 15-year period.


104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

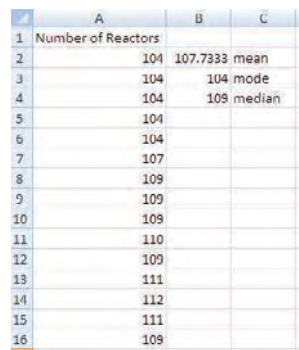
1. On an Excel worksheet enter the numbers in cells A2–A16. Enter a label for the variable in cell A1.

On the same worksheet as the data:

2. Compute the mean of the data: key in **=AVERAGE(A2:A16)** in a blank cell.
3. Compute the mode of the data: key in **=MODE(A2:A16)** in a blank cell.
4. Compute the median of the data: key in **=MEDIAN(A2:A16)** in a blank cell.

These and other statistical functions can also be accessed without typing them into the worksheet directly.

1. Select the Formulas tab from the toolbar and select the Insert Function .
2. Select the Statistical category for statistical functions.
3. Scroll to find the appropriate function and click [OK].



	A	B	C
1	Number of Reactors		
2	104	107.7333	mean
3	104	104	mode
4	104	109	median
5	104		
6	104		
7	107		
8	109		
9	109		
10	109		
11	110		
12	109		
13	111		
14	112		
15	111		
16	109		

(Excel reports only the first mode in a bimodal or multimodal distribution.)

## 3–2 Measures of Variation

In statistics, to describe the data set accurately, statisticians must know more than the measures of central tendency. Consider Example 3–15.

**OBJECTIVE 2**

Describe data, using measures of variation, such as the range, variance, and standard deviation.

**EXAMPLE 3-15 Comparison of Outdoor Paint**

A testing lab wishes to test two experimental brands of outdoor paint to see how long each will last before fading. The testing lab makes 6 gallons of each paint to test. Since different chemical agents are added to each group and only six cans are involved, these two groups constitute two small populations. The results (in months) are shown. Find the mean of each group.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

**SOLUTION**

The mean for brand A is

$$\mu = \frac{\sum X}{N} = \frac{210}{6} = 35 \text{ months}$$

The mean for brand B is

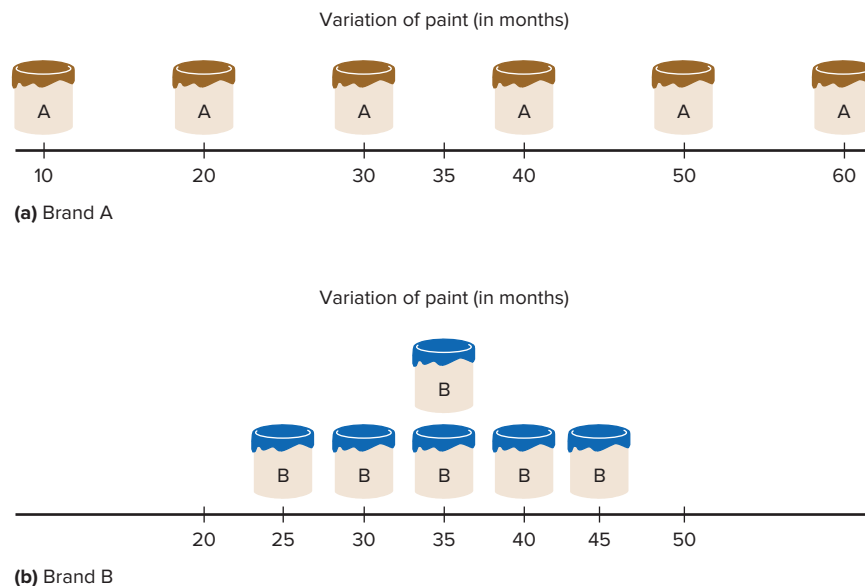
$$\mu = \frac{\sum X}{N} = \frac{210}{6} = 35 \text{ months}$$

Since the means are equal in Example 3-15, you might conclude that both brands of paint last equally well. However, when the data sets are examined graphically, a somewhat different conclusion might be drawn. See Figure 3-2.

As Figure 3-2 shows, even though the means are the same for both brands, the spread, or variation, is quite different. Figure 3-2 shows that brand B performs more consistently; it is less variable. For the spread or variability of a data set, three measures are commonly used: *range*, *variance*, and *standard deviation*. Each measure will be discussed in this section.

**FIGURE 3-2**

Examining Data Sets Graphically



## Range

The range is the simplest of the three measures and is defined now.

The **range** is the highest value minus the lowest value. The symbol  $R$  is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

### EXAMPLE 3–16 Comparison of Outdoor Paint

Find the ranges for the paints in Example 3–15.

#### SOLUTION

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range.

One extremely high or one extremely low data value can affect the range markedly, as shown in Example 3–17.

### EXAMPLE 3–17 Top-Grossing Movies

The data show a sample of the top-grossing movies in millions of dollars for a recent year. Find the range.

\$409    \$386    \$150    \$117    \$73    \$70

Source: *The World Almanac and Book of Facts*.

#### SOLUTION

Range  $R = \$409 - \$70 = \$339$  million

The range for these data is quite large since it depends on the highest data value and the lowest data value. To have a more meaningful statistic to measure the variability, statisticians use measures called the *variance* and *standard deviation*.

## Population Variance and Standard Deviation

Before these measures can be defined, it is necessary to know what *data variation* means. It is based on the difference or distance each data value is from the mean. This difference or distance is called a *deviation*. In the outdoor paint example, the mean for brand A paint is  $\mu = 35$  months, and for a specific can, say, the can that lasted for 50 months, the deviation is  $X - \mu$  or  $50 - 35 = 15$ . Hence, the deviation for that data value is 15 months. If you find the sum of the deviations for all data values about the mean (without rounding), this sum will always be zero. That is,  $\Sigma(X - \mu) = 0$ . (You can see this if you sum all the deviations for the paint example.)

To eliminate this problem, we sum the squares, that is,  $\Sigma(X - \mu)^2$  and find the mean of these squares by dividing by  $N$  (the total number of data values), symbolically  $\Sigma(X - \mu)^2/N$ . This measure is called the *population variance* and is symbolized by  $\sigma^2$ , where  $\sigma$  is the symbol for Greek lowercase letter sigma.

Since this measure ( $\sigma^2$ ) is in square units and the data are in regular units, statisticians take the square root of the variance and call it the *standard deviation*.

Formally defined,

The **population variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is  $\sigma^2$  ( $\sigma$  is the Greek lowercase letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where  $X$  = individual value

$\mu$  = population mean

$N$  = population size

The **population standard deviation** is the square root of the variance. The symbol for the population standard deviation is  $\sigma$ .

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

To find the variance and standard deviation for a data set, the following Procedure Table can be used.

#### Procedure Table

##### Finding the Population Variance and Population Standard Deviation

**Step 1** Find the mean for the data.

$$\mu = \frac{\Sigma X}{N}$$

**Step 2** Find the deviation for each data value.

$$X - \mu$$

**Step 3** Square each of the deviations.

$$(X - \mu)^2$$

**Step 4** Find the sum of the squares.

$$\Sigma(X - \mu)^2$$

**Step 5** Divide by  $N$  to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

**Step 6** Take the square root of the variance to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

#### Interesting Fact

The average American drives about 10,000 miles a year.

**Rounding Rule for the Standard Deviation** The rounding rule for the standard deviation is the same as that for the mean. The final answer should be rounded to one more decimal place than that of the original data.



**EXAMPLE 3–18** Comparison of Outdoor Paint

Find the variance and standard deviation for the data set for brand A paint in Example 3–15. The number of months brand A lasted before fading was

10, 60, 50, 30, 40, 20

**SOLUTION**

**Step 1** Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

**Step 2** Subtract the mean from each data value  $(X - \mu)$ .

$$\begin{array}{lll} 10 - 35 = -25 & 50 - 35 = 15 & 40 - 35 = 5 \\ 60 - 35 = 25 & 30 - 35 = -5 & 20 - 35 = -15 \end{array}$$

**Step 3** Square each result  $(X - \mu)^2$ .

$$\begin{array}{lll} (-25)^2 = 625 & (15)^2 = 225 & (5)^2 = 25 \\ (25)^2 = 625 & (-5)^2 = 25 & (-15)^2 = 225 \end{array}$$

**Step 4** Find the sum of the squares  $\Sigma(X - \mu)^2$ .

$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

**Step 5** Divide the sum by  $N$  to get the variance  $\frac{\Sigma(X - \mu)^2}{N}$ .

$$\text{Variance} = 1750 \div 6 \approx 291.7$$

**Step 6** Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals  $\sqrt{291.7}$ , or 17.1. It is helpful to make a table.

A Values $X$	B $X - \mu$	C $(X - \mu)^2$
10	-25	625
60	25	625
50	15	225
30	-5	25
40	5	25
20	-15	225
		1750

Column A contains the raw data  $X$ . Column B contains the differences  $X - \mu$  obtained in step 2. Column C contains the squares of the differences obtained in step 3.

**Historical Note**

Karl Pearson in 1892 and 1893 introduced the statistical concepts of the range and standard deviation.

The preceding computational procedure reveals several things. First, the square root of the variance gives the standard deviation; and vice versa, squaring the standard deviation gives the variance. Second, the variance is actually the average of the square of the distance that each value is from the mean. Therefore, if the values are near the mean, the variance will be small. In contrast, if the values are far from the mean, the variance will be large.

You might wonder why the squared distances are used instead of the actual distances. As previously stated, the reason is that the sum of the distances will always be zero. To verify this result for a specific case, add the values in column B of the table in Example 3–18. When each value is squared, the negative signs are eliminated.

Finally, why is it necessary to take the square root? Again, the reason is that since the distances were squared, the units of the resultant numbers are the squares of the units of

the original raw data. Finding the square root of the variance puts the standard deviation in the same units as the raw data.

When you are finding the square root, always use its positive value, since the variance and standard deviation of a data set can never be negative.

### EXAMPLE 3-19 Comparison of Outdoor Paint

Find the variance and standard deviation for brand B paint data in Example 3-15. The months brand B lasted before fading were

35, 45, 30, 35, 40, 25

#### SOLUTION

**Step 1** Find the mean.

$$\mu = \frac{\sum X}{N} = \frac{35 + 45 + 30 + 35 + 40 + 25}{6} = \frac{210}{6} = 35$$

**Step 2** Subtract the mean from each value, and place the result in column B of the table.

$$\begin{array}{lll} 35 - 35 = 0 & 45 - 35 = 10 & 30 - 35 = -5 \\ 35 - 35 = 0 & 40 - 35 = 5 & 25 - 35 = -10 \end{array}$$

**Step 3** Square each result and place the squares in column C of the table.

A $X$	B $X - \mu$	C $(X - \mu)^2$
35	0	0
45	10	100
30	-5	25
35	0	0
40	5	25
25	-10	100

**Step 4** Find the sum of the squares in column C.

$$\sum (X - \mu)^2 = 0 + 100 + 25 + 0 + 25 + 100 = 250$$

**Step 5** Divide the sum by  $N$  to get the variance.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N} = \frac{250}{6} = 41.7$$

**Step 6** Take the square root to get the standard deviation.

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}} = \sqrt{41.7} \approx 6.5$$

Hence, the standard deviation is 6.5.

Since the standard deviation of brand A is 17.1 (see Example 3-18) and the standard deviation of brand B is 6.5, the data are more variable for brand A. *In summary, when the means are equal, the larger the variance or standard deviation is, the more variable the data are.*

### Sample Variance and Standard Deviation

When computing the variance for a sample, one might expect the following expression to be used:

$$\frac{\sum (X - \bar{X})^2}{n}$$

where  $\bar{X}$  is the sample mean and  $n$  is the sample size. *This formula is not usually used, however, since in most cases the purpose of calculating the statistic is to estimate the*

corresponding parameter. For example, the sample mean  $\bar{X}$  is used to estimate the population mean  $\mu$ . The expression

$$\frac{\Sigma(X - \bar{X})^2}{n}$$

does not give the best estimate of the population variance because when the population is large and the sample is small (usually less than 30), the variance computed by this formula usually underestimates the population variance. Therefore, instead of dividing by  $n$ , find the variance of the sample by dividing by  $n - 1$ , giving a slightly larger value and an *unbiased* estimate of the population variance.

#### Formula for the Sample Variance

The formula for the sample variance (denoted by  $s^2$ ) is

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1}$$

where  $X$  = individual value

$\bar{X}$  = sample mean

$n$  = sample size

To find the standard deviation of a sample, you must take the square root of the sample variance, which was found by using the preceding formula.

#### Formula for the Sample Standard Deviation

The formula for the sample standard deviation, denoted by  $s$ , is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}}$$

where  $X$  = individual value

$\bar{X}$  = sample mean

$n$  = sample size

The procedure for finding the sample variance and the sample standard deviation is the same as the procedure for finding the population variance and the population standard deviation *except* the sum of the squares is divided by  $n - 1$  (sample size minus 1) instead of  $N$  (population size). Refer to the previous Procedure Table if necessary. The next example shows these steps.

#### EXAMPLE 3-20 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and the sample standard deviation.

9      10      14      7      8      3

Source: Pennsylvania School Board Association.

#### SOLUTION

**Step 1** Find the mean of the data values.

$$\bar{X} = \frac{\Sigma X}{n} = \frac{9 + 10 + 14 + 7 + 8 + 3}{6} = \frac{51}{6} = 8.5$$

**Step 2** Find the deviation for each data value ( $X - \bar{X}$ ).

$$\begin{array}{lll} 9 - 8.5 = 0.5 & 10 - 8.5 = 1.5 & 14 - 8.5 = 5.5 \\ 7 - 8.5 = -1.5 & 8 - 8.5 = -0.5 & 3 - 8.5 = -5.5 \end{array}$$

**Step 3** Square each of the deviations  $(X - \bar{X})^2$ .

$$\begin{array}{lll} (0.5)^2 = 0.25 & (1.5)^2 = 2.25 & (5.5)^2 = 30.25 \\ (-1.5)^2 = 2.25 & (-0.5)^2 = 0.25 & (-5.5)^2 = 30.25 \end{array}$$

**Step 4** Find the sum of the squares.

$$\Sigma(X - \bar{X})^2 = 0.25 + 2.25 + 30.25 + 2.25 + 0.25 + 30.25 = 65.5$$

**Step 5** Divide by  $n - 1$  to get the variance.

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{n - 1} = \frac{65.5}{6 - 1} = \frac{65.5}{5} = 13.1$$

**Step 6** Take the square root of the variance to get the standard deviation.

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{n - 1}} = \sqrt{13.1} \approx 3.6 \text{ (rounded)}$$

Here the sample variance is 13.1, and the sample standard deviation is 3.6.

Shortcut formulas for computing the variance and standard deviation are presented next and will be used in the remainder of the chapter and in the exercises. These formulas are mathematically equivalent to the preceding formulas and do not involve using the mean. They save time when repeated subtracting and squaring occur in the original formulas. They are also more accurate when the mean has been rounded.

#### Shortcut or Computational Formulas for $s^2$ and $s$

The shortcut formulas for computing the variance and standard deviation for data obtained from samples are as follows.

Variance	Standard deviation
$s^2 = \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n - 1)}$	$s = \sqrt{\frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n - 1)}}$

*Note that  $\Sigma X^2$  is not the same as  $(\Sigma X)^2$ . The notation  $\Sigma X^2$  means to square the values first, then sum;  $(\Sigma X)^2$  means to sum the values first, then square the sum.*

Example 3–21 explains how to use the shortcut formulas.

#### EXAMPLE 3–21 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and sample standard deviation.

9, 10, 14, 7, 8, 3

#### SOLUTION

**Step 1** Find the sum of the values:

$$\Sigma X = 9 + 10 + 14 + 7 + 8 + 3 = 51$$

**Step 2** Square each value and find the sum:

$$\Sigma X^2 = 9^2 + 10^2 + 14^2 + 7^2 + 8^2 + 3^2 = 499$$

**Step 3** Substitute in the formula and solve:

$$\begin{aligned} s^2 &= \frac{n(\Sigma X^2) - (\Sigma X)^2}{n(n - 1)} \\ &= \frac{6(499) - 51^2}{6(6 - 1)} \end{aligned}$$

$$\begin{aligned}
 &= \frac{2994 - 2601}{6(5)} \\
 &= \frac{393}{30} \\
 &= 13.1
 \end{aligned}$$

The variance is 13.1.

$$s = \sqrt{13.1} \approx 3.6 \text{ (rounded)}$$

Hence, the sample variance is 13.1, and the sample standard deviation is 3.6. Notice that these are the same results as the results in Example 3-20.

### Variance and Standard Deviation for Grouped Data

The procedure for finding the variance and standard deviation for grouped data is similar to that for finding the mean for grouped data, and it uses the midpoints of each class.

This procedure uses the shortcut formula, and  $X_m$  is the symbol for the class midpoint.

#### Shortcut or Computational Formula for $s^2$ and $s$ for Grouped Data

Sample variance:

$$s^2 = \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}$$

Sample standard deviation

$$s = \sqrt{\frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}}$$

where  $X_m$  is the midpoint of each class and  $f$  is the frequency of each class.

The steps for finding the sample variance and sample standard deviation for grouped data are summarized in this Procedure Table.

#### Procedure Table

##### Finding the Sample Variance and Standard Deviation for Grouped Data

**Step 1** Make a table as shown, and find the midpoint of each class.

A Class	B Frequency	C Midpoint	D $f \cdot X_m$	E $f \cdot X_m^2$
------------	----------------	---------------	--------------------	----------------------

**Step 2** Multiply the frequency by the midpoint for each class, and place the products in column D.

**Step 3** Multiply the frequency by the square of the midpoint, and place the products in column E.

**Step 4** Find the sums of columns B, D, and E. (The sum of column B is  $n$ . The sum of column D is  $\sum f \cdot X_m$ . The sum of column E is  $\sum f \cdot X_m^2$ .)

**Step 5** Substitute in the formula and solve to get the variance.

$$s^2 = \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}$$

**Step 6** Take the square root to get the standard deviation.

**EXAMPLE 3-22 Miles Run per Week**

Find the sample variance and the sample standard deviation for the frequency distribution of the data shown. The data represent the number of miles that 20 runners ran during one week.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

**SOLUTION****Unusual Stat**

At birth men outnumber women by 2%. By age 25, the number of men living is about equal to the number of women living. By age 65, there are 14% more women living than men.

**Step 1** Make a table as shown, and find the midpoint of each class.

A Class	B Frequency $f$	C Midpoint $X_m$	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8		
10.5–15.5	2	13		
15.5–20.5	3	18		
20.5–25.5	5	23		
25.5–30.5	4	28		
30.5–35.5	3	33		
35.5–40.5	2	38		

**Step 2** Multiply the frequency by the midpoint for each class, and place the products in column D.

$$1 \cdot 8 = 8 \quad 2 \cdot 13 = 26 \quad \dots \quad 2 \cdot 38 = 76$$

**Step 3** Multiply the frequency by the square of the midpoint, and place the products in column E.

$$1 \cdot 8^2 = 64 \quad 2 \cdot 13^2 = 338 \quad \dots \quad 2 \cdot 38^2 = 2888$$

**Step 4** Find the sums of columns B, D, and E. The sum of column B is  $n$ , the sum of column D is  $\Sigma f \cdot X_m$ , and the sum of column E is  $\Sigma f \cdot X_m^2$ . The completed table is shown.

A Class	B Frequency $f$	C Midpoint $X_m$	D $f \cdot X_m$	E $f \cdot X_m^2$
5.5–10.5	1	8	8	64
10.5–15.5	2	13	26	338
15.5–20.5	3	18	54	972
20.5–25.5	5	23	115	2,645
25.5–30.5	4	28	112	3,136
30.5–35.5	3	33	99	3,267
35.5–40.5	<u>2</u>	38	<u>76</u>	<u>2,888</u>
	$n = 20$		$\Sigma f \cdot X_m = 490$	$\Sigma f \cdot X_m^2 = 13,310$



**Step 5** Substitute in the formula and solve for  $s^2$  to get the variance.

$$\begin{aligned}s^2 &= \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)} \\&= \frac{20(13,310) - 490^2}{20(20-1)} \\&= \frac{266,200 - 240,100}{20(19)} \\&= \frac{26,100}{380} \\&\approx 68.7\end{aligned}$$

**Step 6** Take the square root to get the standard deviation.

$$s \approx \sqrt{68.7} \approx 8.3$$

Be sure to use the number found in the sum of column B (i.e., the sum of the frequencies) for  $n$ . Do not use the number of classes.

The three measures of variation are summarized in Table 3-2.

### Unusual Stat

The average number of times that a man cries in a month is 1.4.

**TABLE 3-2 Summary of Measures of Variation**

Measure	Definition	Symbol(s)
Range	Distance between highest value and lowest value	$R$
Variance	Average of the squares of the distance that each value is from the mean	$\sigma^2, s^2$
Standard deviation	Square root of the variance	$\sigma, s$

### Uses of the Variance and Standard Deviation

1. As previously stated, variances and standard deviations can be used to determine the spread of the data. If the variance or standard deviation is large, the data are more dispersed. This information is useful in comparing two (or more) data sets to determine which is more (most) variable.
2. The measures of variance and standard deviation are used to determine the consistency of a variable. For example, in the manufacture of fittings, such as nuts and bolts, the variation in the diameters must be small, or else the parts will not fit together.
3. The variance and standard deviation are used to determine the number of data values that fall within a specified interval in a distribution. For example, Chebyshev's theorem (explained later) shows that, for any distribution, at least 75% of the data values will fall within 2 standard deviations of the mean.
4. Finally, the variance and standard deviation are used quite often in inferential statistics. These uses will be shown in later chapters of this textbook.

### Historical Note

Karl Pearson devised the coefficient of variation to compare the deviations of two different groups such as the heights of men and women.

### Coefficient of Variation

Whenever two samples have the same units of measure, the variance and standard deviation for each can be compared directly. For example, suppose an automobile dealer wanted to compare the standard deviation of miles driven for the cars she received as trade-ins on new cars. She found that for a specific year, the standard deviation for Buicks was 422 miles and the standard deviation for Cadillacs was 350 miles. She could say that the variation in mileage was greater in the Buicks. But what if a manager wanted to compare the standard deviations of two different variables, such as the

number of sales per salesperson over a 3-month period and the commissions made by these salespeople?

A statistic that allows you to compare standard deviations when the units are different, as in this example, is called the *coefficient of variation*.

The **coefficient of variation**, denoted by  $CVar$ , is the standard deviation divided by the mean. The result is expressed as a percentage.

**For samples,**

$$CVar = \frac{s}{\bar{X}} \cdot 100$$

**For populations,**

$$CVar = \frac{\sigma}{\mu} \cdot 100$$

### EXAMPLE 3-23 Sales of Automobiles

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is \$5225, and the standard deviation is \$773. Compare the variations of the two.

#### SOLUTION

The coefficients of variation are

$$CVar = \frac{s}{\bar{X}} \cdot 100 = \frac{5}{87} \cdot 100 \approx 5.7\% \quad \text{sales}$$

$$CVar = \frac{773}{5225} \cdot 100 \approx 14.8\% \quad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

### EXAMPLE 3-24 Roller Coasters

The mean speed for the five fastest wooden roller coasters is 69.16 miles per hour, and the variance is 2.76. The mean height for the five tallest roller coasters is 177.80 feet, and the variance is 157.70. Compare the variations of the two data sets.

Source: Ultimate RollerCoaster.com

#### SOLUTION

$$CVar = \frac{\sqrt{2.76}}{69.16} \cdot 100 \approx 2.4\%$$

$$CVar = \frac{\sqrt{157.70}}{177.80} \cdot 100 \approx 7.1\%$$

The variation in the speeds is slightly larger than the variation in the heights of the roller coasters.

## Range Rule of Thumb

The range can be used to approximate the standard deviation. The approximation is called the **range rule of thumb**.

#### The Range Rule of Thumb

A rough estimate of the standard deviation is

$$s \approx \frac{\text{range}}{4}$$

In other words, if the range is divided by 4, an approximate value for the standard deviation is obtained. For example, the standard deviation for the data set 5, 8, 8, 9, 10, 12, and 13 is 2.7, and the range is  $13 - 5 = 8$ . The range rule of thumb is  $s \approx 2$ . The range rule of thumb in this case underestimates the standard deviation somewhat; however, it is in the ballpark.

A note of caution should be mentioned here. The range rule of thumb is only an *approximation* and should be used when the distribution of data values is unimodal and roughly symmetric.

The range rule of thumb can be used to estimate the largest and smallest data values of a data set. The smallest data value will be approximately 2 standard deviations below the mean, and the largest data value will be approximately 2 standard deviations above the mean of the data set. The mean for the previous data set is 9.3; hence,

$$\text{Smallest data value} = \bar{X} - 2s = 9.3 - 2(2.7) = 3.9$$

$$\text{Largest data value} = \bar{X} + 2s = 9.3 + 2(2.7) = 14.7$$

Notice that the smallest data value was 5, and the largest data value was 13. Again, these are rough approximations. For many data sets, almost all data values will fall within 2 standard deviations of the mean. Better approximations can be obtained by using Chebyshev's theorem and the empirical rule. These are explained next.

### Chebyshev's Theorem

As stated previously, the variance and standard deviation of a variable can be used to determine the spread, or dispersion, of a variable. That is, the larger the variance or standard deviation, the more the data values are dispersed. For example, if two variables measured in the same units have the same mean, say, 70, and the first variable has a standard deviation of 1.5 while the second variable has a standard deviation of 10, then the data for the second variable will be more spread out than the data for the first variable. *Chebyshev's theorem*, developed by the Russian mathematician Chebyshev (1821–1894), specifies the proportions of the spread in terms of the standard deviation.

**Chebyshev's theorem** The proportion of values from a data set that will fall within  $k$  standard deviations of the mean will be at least  $1 - 1/k^2$ , where  $k$  is a number greater than 1 ( $k$  is not necessarily an integer).

This theorem states that at least three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean of the data set. This result is found by substituting  $k = 2$  in the expression

$$1 - \frac{1}{k^2} \quad \text{or} \quad 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 75\%$$

For the example in which variable 1 has a mean of 70 and a standard deviation of 1.5, at least three-fourths, or 75%, of the data values fall between 67 and 73. These values are found by adding 2 standard deviations to the mean and subtracting 2 standard deviations from the mean, as shown:

$$70 + 2(1.5) = 70 + 3 = 73$$

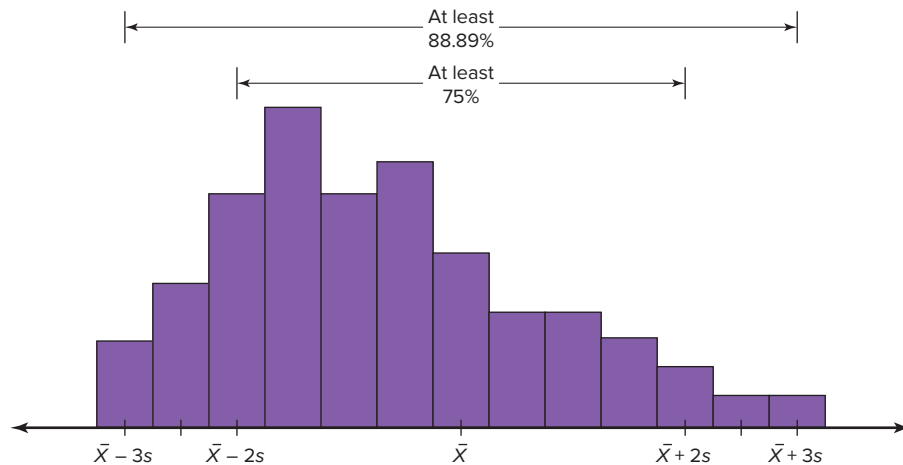
and

$$70 - 2(1.5) = 70 - 3 = 67$$

For variable 2, at least three-fourths, or 75%, of the data values fall between 50 and 90. Again, these values are found by adding and subtracting, respectively, 2 standard deviations to and from the mean.

$$70 + 2(10) = 70 + 20 = 90$$

**FIGURE 3-3**  
Chebyshev's Theorem



and

$$70 - 2(10) = 70 - 20 = 50$$

Furthermore, the theorem states that at least eight-ninths, or 88.89%, of the data values will fall within 3 standard deviations of the mean. This result is found by letting  $k = 3$  and substituting in the expression.

$$1 - \frac{1}{k^2} \quad \text{or} \quad 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 88.89\%$$

For variable 1, at least eight-ninths, or 88.89%, of the data values fall between 65.5 and 74.5, since

$$70 + 3(1.5) = 70 + 4.5 = 74.5$$

and

$$70 - 3(1.5) = 70 - 4.5 = 65.5$$

For variable 2, at least eight-ninths, or 88.89%, of the data values fall between 40 and 100.

In summary, then, Chebyshev's theorem states

- At least three-fourths, or 75%, of all data values fall within 2 standard deviations of the mean.
- At least eight-ninths, or 89%, of all data values fall within 3 standard deviations of the mean.

This theorem can be applied to any distribution regardless of its shape (see Figure 3-3). Examples 3-25 and 3-26 illustrate the application of Chebyshev's theorem.

### EXAMPLE 3-25 Prices of Homes

The mean price of houses in a certain neighborhood is \$50,000, and the standard deviation is \$10,000. Find the price range for which at least 75% of the houses will sell.

#### SOLUTION

Chebyshev's theorem states that three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean. Thus,

$$\$50,000 + 2(\$10,000) = \$50,000 + \$20,000 = \$70,000$$

and

$$\$50,000 - 2(\$10,000) = \$50,000 - \$20,000 = \$30,000$$

Hence, at least 75% of all homes sold in the area will have a price range from \$30,000 to \$70,000.

Chebyshev's theorem can be used to find the minimum percentage of data values that will fall between any two given values. The procedure is shown in Example 3-26.

### EXAMPLE 3-26 Travel Allowances

A survey of local companies found that the mean amount of travel allowance for couriers was \$0.25 per mile. The standard deviation was \$0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

#### SOLUTION

**Step 1** Subtract the mean from the larger value.

$$\$0.30 - \$0.25 = \$0.05$$

**Step 2** Divide the difference by the standard deviation to get  $k$ .

$$k = \frac{0.05}{0.02} = 2.5$$

**Step 3** Use Chebyshev's theorem to find the percentage.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = 1 - 0.16 = 0.84 \quad \text{or} \quad 84\%$$

Hence, at least 84% of the data values will fall between \$0.20 and \$0.30.

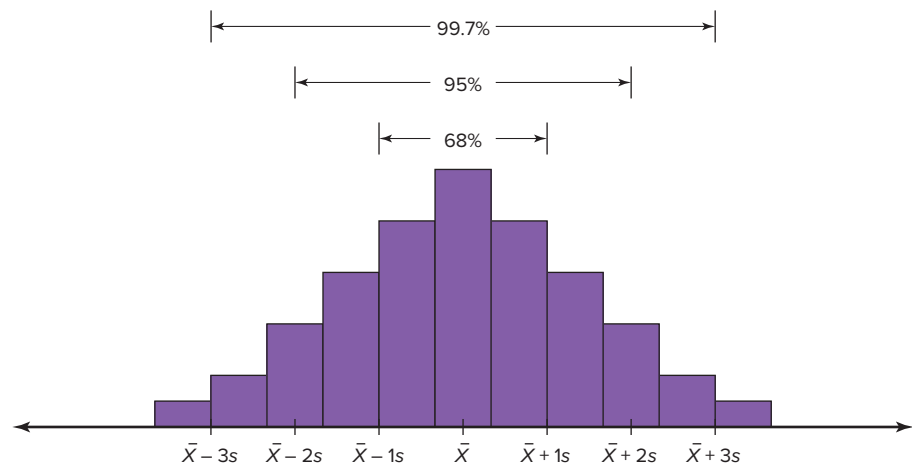
### The Empirical (Normal) Rule

Chebyshev's theorem applies to any distribution regardless of its shape. However, when a distribution is *bell-shaped* (or what is called *normal*), the following statements, which make up the **empirical rule**, are true.

- Approximately 68% of the data values will fall within 1 standard deviation of the mean.
- Approximately 95% of the data values will fall within 2 standard deviations of the mean.
- Approximately 99.7% of the data values will fall within 3 standard deviations of the mean.

For example, suppose that the scores on a national achievement exam have a mean of 480 and a standard deviation of 90. If these scores are normally distributed, then approximately 68% will fall between 390 and 570 ( $480 + 90 = 570$  and  $480 - 90 = 390$ ). Approximately 95% of the scores will fall between 300 and 660 ( $480 + 2 \cdot 90 = 660$  and  $480 - 2 \cdot 90 = 300$ ). Approximately 99.7% will fall between 210 and 750 ( $480 + 3 \cdot 90 = 750$  and  $480 - 3 \cdot 90 = 210$ ). See Figure 3-4. (The empirical rule is explained in greater detail in Chapter 6.)

**FIGURE 3-4**  
The Empirical Rule



Because the empirical rule requires that the distribution be approximately bell-shaped, the results are more accurate than those of Chebyshev's theorem, which applies to all distributions.

### Linear Transformation of Data

In statistics, sometimes it is necessary to transform the data values into other data values. For example, if you are using temperature values collected from Canada, these values will be given using the Celsius temperature scale. If the study is to be used in the United States, you might want to change the data values to the Fahrenheit temperature scale. This change is called the *linear transformation of the data*. The question then arises, How does a linear transformation of the data values affect the mean and standard deviation of the data values?

Let's look at an example. Suppose you own a small store with five employees. Their hourly salaries are

\$10      \$13      \$10      \$11      \$16

The mean of the salaries is  $\bar{X} = \$12$ , and the standard deviation is 2.550. Suppose you decide after a profitable year to give each employee a raise of \$1.00 per hour. The new salaries would be

\$11      \$14      \$11      \$12      \$17

The mean of the new salaries is  $\bar{X} = \$13$ , and the standard deviation of the new salaries is 2.550. Notice that the value of the mean increases by the amount that was added to each data value, and the standard deviation does not change.

Suppose that the five employees worked the number of hours per week shown here.

15      12      18      20      10

The mean for the number of hours for the original data is  $\bar{X} = 15$ , and the standard deviation for the number of hours is 4.123. You next decide to double the amount of each employee's hours for December. How does this affect the mean and standard deviation of the variable?

If each data value is multiplied by 2, the new data set is

30      24      36      40      20

The mean of the new data set is  $\bar{X} = 30$ , and the standard deviation is 8.246. The values of the mean and standard deviation double.

Hence, when each data value is multiplied by a constant, the mean of the new data set will be equal to the constant times the mean of the original set, and the standard deviation of the new data set will be equal to the absolute value (positive value) of the constant times the standard deviation of the original data set.

### Applying the Concepts 3-2

#### Blood Pressure

The table lists means and standard deviations. The mean is the number before the plus/minus, and the standard deviation is the number after the plus/minus. The results are from a study attempting to find the average blood pressure of older adults. Use the results to answer the questions.

	Normotensive		Hypertensive	
	Men ( <i>n</i> = 1200)	Women ( <i>n</i> = 1400)	Men ( <i>n</i> = 1100)	Women ( <i>n</i> = 1300)
Age	55 ± 10	55 ± 10	60 ± 10	64 ± 10
Blood pressure (mm Hg)				
Systolic	123 ± 9	121 ± 11	153 ± 17	156 ± 20
Diastolic	78 ± 7	76 ± 7	91 ± 10	88 ± 10



1. Apply Chebyshev's theorem to the systolic blood pressure of normotensive men. At least how many of the men in the study fall within 1 standard deviation of the mean?
2. At least how many of those men in the study fall within 2 standard deviations of the mean?

Assume that blood pressure is normally distributed among older adults. Answer the following questions, using the empirical rule instead of Chebyshev's theorem.

3. What are the ranges for the diastolic blood pressure (normotensive and hypertensive) of older women?
4. Do the normotensive, male, systolic blood pressure ranges overlap with the hypertensive, male, systolic blood pressure ranges?

See page 184 for the answers.

## Exercises 3-2

1. What is the relationship between the variance and the standard deviation?
2. Why might the range *not* be the best estimate of variability?
3. What are the symbols used to represent the population variance and standard deviation?
4. What are the symbols used to represent the sample variance and standard deviation?
5. Why is the unbiased estimator of variance used?
6. The three data sets have the same mean and range, but is the variation the same? Prove your answer by computing the standard deviation. Assume the data were obtained from samples.
  - a. 5, 7, 9, 11, 13, 15, 17
  - b. 5, 6, 7, 11, 15, 16, 17
  - c. 5, 5, 5, 11, 17, 17, 17

7. **Traveler Spending** The data show the traveler spending in billions of dollars for a recent year for a sample of the states. Find the range, variance, and standard deviation for the data.

20.1   33.5   21.7   58.4   23.2   110.8   30.9  
24.0   74.8   60.0

Source: U.S. Travel Agency.

8. **Cigarette Taxes** The increases (in cents) in cigarette taxes for 17 states in a 6-month period are  
60, 20, 40, 40, 45, 12, 34, 51, 30, 70, 42, 31, 69, 32, 8, 18, 50  
Find the range, variance, and standard deviation for the data. Use the range rule of thumb to estimate the standard deviation. Compare the estimate to the actual standard deviation.

Source: Federation of Tax Administrators.

9. **Prices of Silver and Tin** The following data show the price of silver and the price of tin over a recent 9-year

period. Find the range, variance, and standard deviation. Which data set is more variable?

Silver	Tin
23.80	13.40
31.21	12.83
35.24	15.75
20.20	12.40
14.69	8.37
15.00	11.29
13.41	8.99
11.57	5.65
7.34	4.83

Source: Department of the Interior.

10. **Size of U.S. States** The total surface area (in square miles) for each of six selected eastern states is listed here.

28,995	PA	37,534	FL
31,361	NY	27,087	VA
20,966	ME	37,741	GA

The total surface area for each of six selected western states is listed (in square miles).

72,964	AZ	70,763	NV
101,510	CA	62,161	OR
66,625	CO	54,339	UT

Find the standard deviation for each data set. Which set is more variable?

Source: New York Times Almanac.

11. **Multiple Births** The numbers of various multiple births in the United States for the past 10 years are listed. Find the range, variance, and standard deviation of the data sets. Which set of data is the most variable?

Triplets			Quadruplets			Quintuplets		
5877	7110	5937	345	468	369	46	85	91
6898	6118	6885	434	355	501	69	67	85
6208	6742	6750	418	506	439	68	77	86
	6742			512			67	

Source: World Almanac 2012.

12. **Starting Teachers' Salaries** Starting teachers' salaries (in equivalent U.S. dollars) for upper secondary education in selected countries are listed. Find the range, variance, and standard deviation for the data. Which set of data is

more variable? (The U.S. average starting salary at this time was \$29,641.)

Europe		Asia	
Sweden	\$48,704	Korea	\$26,852
Germany	41,441	Japan	23,493
Spain	32,679	India	18,247
Finland	32,136	Malaysia	13,647
Denmark	30,384	Philippines	9,857
Netherlands	29,326	Thailand	5,862
Scotland	27,789		

Source: World Almanac.

- 13. Ages of U.S. Astronaut Candidates** The average age of U.S. astronaut candidates in the past has been 34, but candidates have ranged in age from 26 to 46. Use the range rule of thumb to estimate the standard deviation of the applicants' ages.

Source: www.nasa.gov

- 14. Times Spent in Rush-Hour Traffic** A sample of 12 drivers shows the time that they spent (in minutes) stopped in rush-hour traffic on a specific snowy day last winter. Find the range, variance, and standard deviation for the data.

52	56	53
61	49	51
53	58	53
60	71	58

- 15. Laws Passed** The data show the number of public laws passed by the U.S. Congress for a sample of recent years. Find the range, variance, and standard deviation for the data.
- 283 394 383 580 498 460 377 482

Source: Congressional Record.

- 16. Passenger Vehicle Deaths** The number of people killed in each state from passenger vehicle crashes for a specific year is shown. Find the range, variance, and standard deviation for the data.

778	309	1110	324	705
1067	826	76	205	152
218	492	65	186	712
193	262	452	875	82
730	1185	2707	1279	390
305	123	948	343	602
69	451	951	104	985
155	450	2080	565	875
414	981	2786	82	793
214	130	396	620	797

Source: National Highway Traffic Safety Administration.

- 17. Annual Precipitation Days** The number of annual precipitation days for one-half of the 50 largest U.S. cities is listed below. Find the range, variance, and standard deviation of the data.

135	128	136	78	116	77	111	79	44	97
116	123	88	102	26	82	156	133	107	35
112	98	45	122	125					

- 18.** Use the data from Exercises 7, 15, and 17 (unemployment, prisoners, precipitation days) and compare the standard

deviation with that obtained by the range rule of thumb (R/4.) Comment on the results.

- 19. Pupils Per Teacher** The following frequency distribution shows the average number of pupils per teacher in the 50 states of the United States. Find the variance and standard deviation for the data.

Class limits	Frequency
9–11	2
12–14	20
15–17	18
18–20	7
21–23	2
24–26	1
	<u>50</u>

Source: U.S. Department of Education.

- 20. Automotive Fuel Efficiency** Thirty automobiles were tested for fuel efficiency (in miles per gallon). This frequency distribution was obtained. Find the variance and standard deviation for the data.

Class boundaries	Frequency
7.5–12.5	3
12.5–17.5	5
17.5–22.5	15
22.5–27.5	5
27.5–32.5	2

- 21. Murders in Cities** The data show the number of murders in 25 selected cities. Find the variance and standard deviation for the data.

Class limits	Frequency
34–96	13
97–159	2
160–222	0
223–285	5
286–348	1
349–411	1
412–474	0
475–537	1
538–600	2

- 22. Reaction Times** In a study of reaction times to a specific stimulus, a psychologist recorded these data (in seconds). Find the variance and standard deviation for the data.

Class limits	Frequency
2.1–2.7	12
2.8–3.4	13
3.5–4.1	7
4.2–4.8	5
4.9–5.5	2
5.6–6.2	1

- 23. FM Radio Stations** A random sample of 30 states shows the number of low-power FM radio stations for each state. Find the variance and standard deviation for the data.

Class limits	Frequency
1–9	5
10–18	7
19–27	10
28–36	3
37–45	3
46–54	2

Source: Federal Communications Commission.

- 24. Murder Rates** The data represent the murder rate per 100,000 individuals in a sample of selected cities in the United States. Find the variance and standard deviation for the data.

Class limits	Frequency
5–11	8
12–18	5
19–25	7
26–32	1
33–39	1
40–46	3

Source: FBI and U.S. Census Bureau.

- 25. Waterfall Heights** The frequency distribution shows a sample of the waterfall heights, in feet, of 28 waterfalls. Find the variance and standard deviation for the data.

Class boundaries	Frequency
52.5–185.5	8
185.5–318.5	11
318.5–451.5	2
451.5–584.5	1
584.5–717.5	4
717.5–850.5	2

Source: National Geographic Society.

- 26. Baseball Team Batting Averages** Team batting averages for major league baseball in 2015 are represented below. Find the variance and standard deviation for each league. Compare the results.

NL		AL	
0.242–0.246	3	0.244–0.249	3
0.247–0.251	6	0.250–0.255	6
0.252–0.256	1	0.256–0.261	2
0.257–0.261	11	0.262–0.267	1
0.262–0.266	11	0.268–0.273	3
0.267–0.271	1	0.274–0.279	0

Source: World Almanac.

- 27. Missing Work** The average number of days that construction workers miss per year is 11. The standard deviation is 2.3. The average number of days that factory

workers miss per year is 8 with a standard deviation of 1.8. Which class is more variable in terms of days missed?

- 28. Suspension Bridges** The lengths (in feet) of the main span of the longest suspension bridges in the United States and the rest of the world are shown below. Which set of data is more variable?

United States 4205, 4200, 3800, 3500, 3478, 2800, 2800, 2310  
World 6570, 5538, 5328, 4888, 4626, 4544, 4518, 3970

Source: World Almanac.

- 29. Hospital Emergency Waiting Times** The mean of the waiting times in an emergency room is 80.2 minutes with a standard deviation of 10.5 minutes for people who are admitted for additional treatment. The mean waiting time for patients who are discharged after receiving treatment is 120.6 minutes with a standard deviation of 18.3 minutes. Which times are more variable?

- 30. Ages of Accountants** The average age of the accountants at Three Rivers Corp. is 26 years, with a standard deviation of 6 years; the average salary of the accountants is \$31,000, with a standard deviation of \$4000. Compare the variations of age and income.

- 31.** Using Chebyshev's theorem, solve these problems for a distribution with a mean of 80 and a standard deviation of 10.

- At least what percentage of values will fall between 60 and 100?
- At least what percentage of values will fall between 65 and 95?

- 32.** The mean of a distribution is 20 and the standard deviation is 2. Use Chebyshev's theorem.

- At least what percentage of the values will fall between 10 and 30?
- At least what percentage of the values will fall between 12 and 28?

- 33.** In a distribution of 160 values with a mean of 72, at least 120 fall within the interval 67–77. Approximately what percentage of values should fall in the interval 62–82? Use Chebyshev's theorem.

- 34. Calories in Bagels** The average number of calories in a regular-size bagel is 240. If the standard deviation is 38 calories, find the range in which at least 75% of the data will lie. Use Chebyshev's theorem.

- 35. Time Spent Online** Americans spend an average of 3 hours per day online. If the standard deviation is 32 minutes, find the range in which at least 88.89% of the data will lie. Use Chebyshev's theorem.

Source: www.cs.cmu.edu

- 36. Solid Waste Production** The average college student produces 640 pounds of solid waste each year. If the standard deviation is approximately 85 pounds, within what weight limits will at least 88.89% of all students' garbage lie?  
*Source:* Environmental Sustainability Committee, [www.esc.mtu.edu](http://www.esc.mtu.edu)
- 37. Sale Price of Homes** The average sale price of one-family houses in the United States for January 2016 was \$258,100. Find the range of values in which at least 75% of the sale prices will lie if the standard deviation is \$48,500.  
*Source:* YCharts.com
- 38. Trials to Learn a Maze** The average of the number of trials it took a sample of mice to learn to traverse a maze was 12. The standard deviation was 3. Using Chebyshev's theorem, find the minimum percentage of data values that will fall in the range of 4–20 trials.
- 39. Farm Sizes** The average farm in the United States in 2014 contained 504 acres. The standard deviation is 55.7 acres. Use Chebyshev's theorem to find the minimum percentage of data values that will fall in the range of 364.75 and 643.25 acres.  
*Source:* World Almanac.
- 40. Citrus Fruit Consumption** The average U.S. yearly per capita consumption of citrus fruit is 26.8 pounds. Suppose that the distribution of fruit amounts consumed is bell-shaped with a standard deviation equal to 4.2 pounds. What percentage of Americans would you expect to consume more than 31 pounds of citrus fruit per year?  
*Source:* USDA/Economic Research Service.
- 41. SAT Scores** The national average for mathematics SATs in 2014 was 538. Suppose that the distribution of scores was approximately bell-shaped and that the standard deviation was approximately 48. Within what boundaries

would you expect 68% of the scores to fall? What percentage of scores would be above 634?

- 42. Work Hours for College Faculty** The average full-time faculty member in a postsecondary degree-granting institution works an average of 53 hours per week.
- If we assume the standard deviation is 2.8 hours, what percentage of faculty members work more than 58.6 hours a week?
  - If we assume a bell-shaped distribution, what percentage of faculty members work more than 58.6 hours a week?
- Source:* National Center for Education Statistics.
- 43. Prices of Musical Instruments** The average price of an instrument at a small music store is \$325. The standard deviation of the price is \$52. If the owner decides to raise the price of all the instruments by \$20, what will be the new mean and standard deviation of the prices?
- 44. Hours of Employment** The mean and standard deviation of the number of hours the employees work in the music store per week are, respectively, 18.6 and 3.2 hours. If the owner increases the number of hours each employee works per week by 4 hours, what will be the new mean and standard deviation of the number of hours worked by the employees?
- 45. Price of Pet Fish** The mean price of the fish in a pet shop is \$2.17, and the standard deviation of the price is \$0.55. If the owner decides to triple the prices, what will be the mean and standard deviation of the new prices?
- 46. Bonuses** The mean and standard deviation of the bonuses that the employees of a company received 10 years ago were, respectively, \$2,000 and \$325. Today the amount of the bonuses is 5 times what it was 10 years ago. Find the mean and standard deviation of the new bonuses.

## Extending the Concepts

- 47. Serum Cholesterol Levels** For this data set, find the mean and standard deviation of the variable. The data represent the serum cholesterol levels of 30 individuals. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.
- |     |     |     |     |     |
|-----|-----|-----|-----|-----|
| 211 | 240 | 255 | 219 | 204 |
| 200 | 212 | 193 | 187 | 205 |
| 256 | 203 | 210 | 221 | 249 |
| 231 | 212 | 236 | 204 | 187 |
| 201 | 247 | 206 | 187 | 200 |
| 237 | 227 | 221 | 192 | 196 |
- 48. Ages of Consumers** For this data set, find the mean and standard deviation of the variable. The data represent the ages of 30 customers who ordered a product advertised

on television. Count the number of data values that fall within 2 standard deviations of the mean. Compare this with the number obtained from Chebyshev's theorem. Comment on the answer.

42	44	62	35	20
30	56	20	23	41
55	22	31	27	66
21	18	24	42	25
32	50	31	26	36
39	40	18	36	22

- 49.** Using Chebyshev's theorem, complete the table to find the minimum percentage of data values that fall within  $k$  standard deviations of the mean.

$k$	1.5	2	2.5	3	3.5
Percent					

50. Use this data set: 10, 20, 30, 40, 50
- Find the standard deviation.
  - Add 5 to each value, and then find the standard deviation.
  - Subtract 5 from each value and find the standard deviation.
  - Multiply each value by 5 and find the standard deviation.
  - Divide each value by 5 and find the standard deviation.
  - Generalize the results of parts *b* through *e*.
  - Compare these results with those in Exercise 35 of Exercises 3-1.

51. **Mean Deviation** The mean deviation is found by using this formula:

$$\text{Mean deviation} = \frac{\sum |X - \bar{X}|}{n}$$

where  $X$  = value

$\bar{X}$  = mean

$n$  = number of values

$||$  = absolute value

Find the mean deviation for these data.

5, 9, 10, 11, 11, 12, 15, 18, 20, 22

52. **Pearson Coefficient of Skewness** A measure to determine the skewness of a distribution is called the *Pearson coefficient (PC) of skewness*. The formula is

$$PC = \frac{3(\bar{X} - MD)}{s}$$

The values of the coefficient usually range from  $-3$  to  $+3$ . When the distribution is symmetric, the coefficient is zero; when the distribution is positively skewed, it is positive; and when the distribution is negatively skewed, it is negative.

Using the formula, find the coefficient of skewness for each distribution, and describe the shape of the distribution.

- Mean = 10, median = 8, standard deviation = 3.
  - Mean = 42, median = 45, standard deviation = 4.
  - Mean = 18.6, median = 18.6, standard deviation = 1.5.
  - Mean = 98, median = 97.6, standard deviation = 4.
53. All values of a data set must be within  $s\sqrt{n-1}$  of the mean. If a person collected 25 data values that had a mean of 50 and a standard deviation of 3 and you saw that one data value was 67, what would you conclude?

## Technology

### EXCEL

#### Step by Step

## Step by Step

### Finding Measures of Variation

#### Example XL3-2


Find the sample variance, sample standard deviation, and range of the data from Example 3-20.

9    10    14    7    8    3

- On an Excel worksheet enter the data in cells A2:A7. Enter a label for the variable in cell A1.
- In a blank cell enter = **VAR.S(A2:A7)** for the sample variance.
- In a blank cell enter = **STDEV.S(A2:A7)** for the sample standard deviation.
- For the range, compute the difference between the maximum and the minimum values by entering = **Max(A2:A7)-Min(A2:A7)**.

*Note:* The command for computing the population variance is VAR.P and for the population standard deviation is STDEV.P

These and other statistical functions can also be accessed without typing them into the worksheet directly.

- Select the Formulas tab from the Toolbar and select the Insert Function Icon, .
- Select the Statistical category for statistical functions.
- Scroll to find the appropriate function and click [OK].

	A	B	C	D
1	Strikes			
2	9		Variance	13.1
3	10		Standard Deviation	3.619392214
4	14		Range	11
5	7			
6	8			
7	3			

### 3-3 Measures of Position

#### OBJECTIVE 3

Identify the position of a data value in a data set, using various measures of position, such as percentiles, deciles, and quartiles.

In addition to measures of central tendency and measures of variation, there are measures of position or location. These measures include standard scores, percentiles, deciles, and quartiles. They are used to locate the relative position of a data value in the data set. For example, if a value is located at the 80th percentile, it means that 80% of the values fall below it in the distribution and 20% of the values fall above it. The *median* is the value that corresponds to the 50th percentile, since one-half of the values fall below it and one-half of the values fall above it. This section discusses these measures of position.

#### Standard Scores

There is an old saying, “You can’t compare apples and oranges.” But with the use of statistics, it can be done to some extent. Suppose that a student scored 90 on a music test and 45 on an English exam. Direct comparison of raw scores is impossible, since the exams might not be equivalent in terms of number of questions, value of each question, and so on. However, a comparison of a relative standard similar to both can be made. This comparison uses the mean and standard deviation and is called a *standard score* or *z score*. (We also use *z* scores in later chapters.)

A standard score or *z* score tells how many standard deviations a data value is above or below the mean for a specific distribution of values. If a standard score is zero, then the data value is the same as the mean.

A **z score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \bar{X}}{s}$$

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

For the purpose of this section, it will be assumed that when we find *z* scores, the data were obtained from samples.

#### EXAMPLE 3-27 Test Scores

A student scored 85 on an English test while the mean score of all the students was 76 and the standard deviation was 4. She also scored 42 on a French test where the class mean was 36 and the standard deviation was 3. Compare the relative positions on the two tests.

#### SOLUTION

First find the *z* scores. For the English test

$$z = \frac{X - \bar{X}}{s} = \frac{85 - 76}{4} = 2.25$$

#### Interesting Fact

The average number of faces that a person learns to recognize and remember during his or her lifetime is 10,000.



For the French test

$$z = \frac{X - \bar{X}}{s} = \frac{42 - 36}{3} = 2.00$$

Since the  $z$  score for the English test is higher than the  $z$  score for the French test, her relative position in the English class is higher than her relative position in the French class.

Note that if the  $z$  score is positive, the score is above the mean. If the  $z$  score is 0, the score is the same as the mean. And if the  $z$  score is negative, the score is below the mean.

### EXAMPLE 3-28 Marriage Ages

In a recent study, the mean age at which men get married is said to be 26.4 years with a standard deviation of 2 years. The mean age at which women marry is 23.5 years with a standard deviation of 2.3 years. Find the relative positions for a man who marries at age 24 and a woman who marries at age 22.

#### SOLUTION

$$\text{Man } z = \frac{X - \bar{X}}{s} = \frac{24 - 26.4}{2} = -1.2$$

$$\text{Woman } z = \frac{X - \bar{X}}{s} = \frac{22 - 23.5}{2.3} \approx -0.65$$

In this case, the woman's age at marriage is relatively higher than the man's age at marriage.

When all data for a variable are transformed into  $z$  scores, the resulting distribution will have a mean of 0 and a standard deviation of 1. A  $z$  score, then, is actually the number of standard deviations each value is from the mean for a specific distribution. In Example 3-27, the English test score was 2.25 standard deviations above the mean, while the French test score was 2 standard deviations above the mean. This will be explained in greater detail in Chapter 6.

## Percentiles

Percentiles are position measures used in educational and health-related fields to indicate the position of an individual in a group.

**Percentiles** divide the data set into 100 equal groups.

Percentiles are symbolized by

$$P_1, P_2, P_3, \dots, P_{99}$$

and divide the distribution into 100 groups.



**Interesting Facts**

The highest recorded temperature on earth was 136°F in Libya in 1922. The lowest recorded temperature on earth was -129°F in Antarctica in 1983.

**TABLE 3–3 Percentile Ranks and Scaled Scores on the Test of English as a Foreign Language\***

Scaled score	Section 1: Listening comprehension	Section 2: Structure and written expression	Section 3: Vocabulary and reading comprehension	Total scaled score	Percentile rank
68	99	98			
66	98	96	98	660	99
64	96	94	96	640	97
62	92	90	93	620	94
60	87	84	88	600	89
→58	81	76	81	580	82
56	73	68	72	560	73
54	64	58	61	540	62
52	54	48	50	520	50
50	42	38	40	500	39
48	32	29	30	480	29
46	22	21	23	460	20
44	14	15	16	440	13
42	9	10	11	420	9
40	5	7	8	400	5
38	3	4	5	380	3
36	2	3	3	360	1
34	1	2	2	340	1
32		1	1	320	
30		1	1	300	
Mean	51.5	52.2	51.4	517	Mean
S.D.	7.1	7.9	7.5	68	S.D.

\*Based on the total group of 1,178,193 examinees.

Source: Data from Educational Testing Service.

In many situations, the graphs and tables showing the percentiles for various measures such as test scores, heights, or weights have already been completed. Table 3–3 shows the percentile ranks for scaled scores on the Test of English as a Foreign Language. If a student had a scaled score of 58 for section 1 (listening and comprehension), that student would have a percentile rank of 81. Hence, that student did better than 81% of the students who took section 1 of the exam.

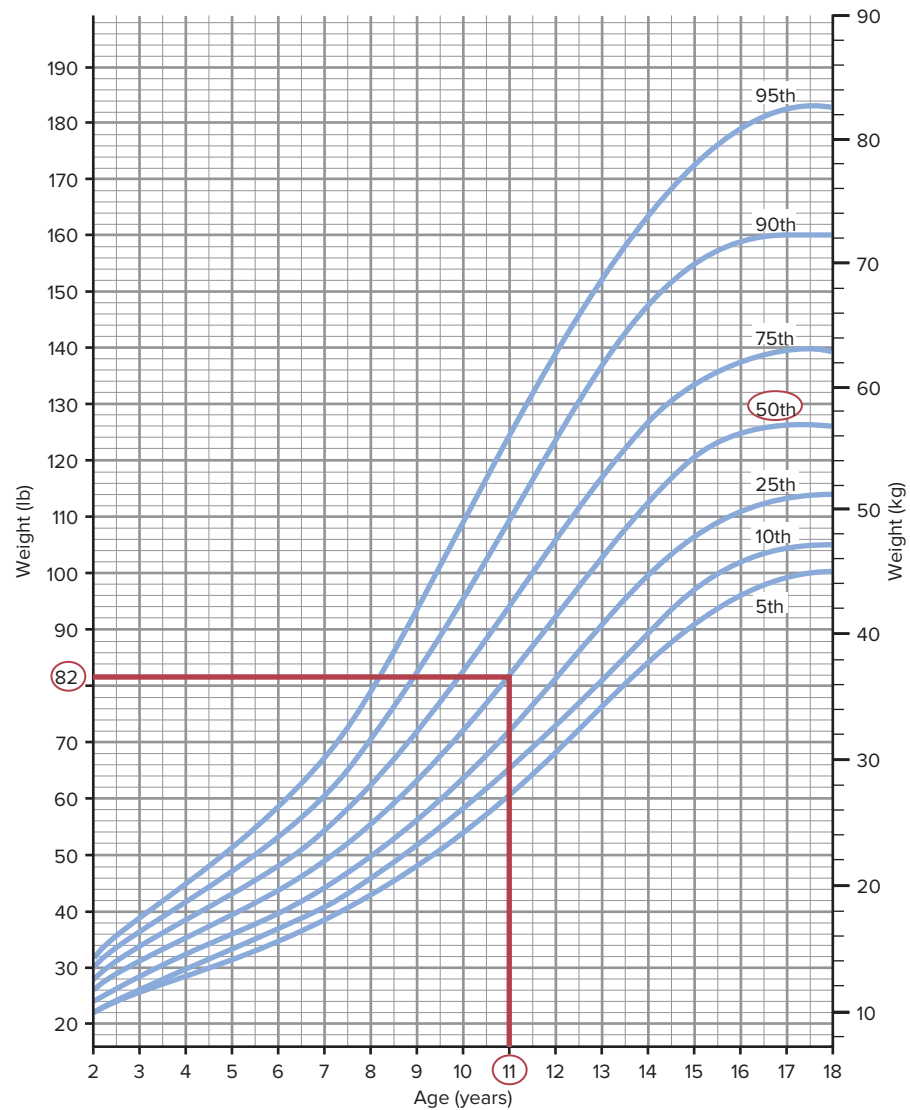
Figure 3–5 shows percentiles in graphical form of weights of girls from ages 2 to 18. To find the percentile rank of an 11-year-old who weighs 82 pounds, start at the 82-pound weight on the left axis and move horizontally to the right. Find 11 on the horizontal axis and move up vertically. The two lines meet at the 50th percentile curved line; hence, an 11-year-old girl who weighs 82 pounds is in the 50th percentile for her age group. If the lines do not meet exactly on one of the curved percentile lines, then the percentile rank must be approximated.

Percentiles are also used to compare an individual's test score with the national norm. For example, tests such as the National Educational Development Test (NEDT) are taken by students in ninth or tenth grade. A student's scores are compared with those of other

**FIGURE 3–5**

Weights of Girls by Age and Percentile Rankings

Source: Centers for Disease Control and Prevention



students locally and nationally by using percentile ranks. A similar test for elementary school students is called the California Achievement Test.

Percentiles are not the same as percentages. That is, if a student gets 72 correct answers out of a possible 100, she obtains a percentage score of 72. There is no indication of her position with respect to the rest of the class. She could have scored the highest, the lowest, or somewhere in between. On the other hand, if a raw score of 72 corresponds to the 64th percentile, then she did better than 64% of the students in her class.

Percentile graphs can be constructed as shown in Example 3–29 and Figure 3–6. Percentile graphs use the same values as the cumulative relative frequency graphs described in Section 2–2, except that the proportions have been converted to percents.

### EXAMPLE 3–29 Systolic Blood Pressure

The frequency distribution for the systolic blood pressure readings (in millimeters of mercury, mm Hg) of 200 randomly selected college students is shown here. Construct a percentile graph.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24		
104.5–119.5	62		
119.5–134.5	72		
134.5–149.5	26		
149.5–164.5	12		
164.5–179.5	4		
	<u>200</u>		

**SOLUTION**

**Step 1** Find the cumulative frequencies and place them in column C.

**Step 2** Find the cumulative percentages and place them in column D. To do this step, use the formula

$$\text{Cumulative \%} = \frac{\text{cumulative frequency}}{n} \cdot 100$$

For the first class,

$$\text{Cumulative \%} = \frac{24}{200} \cdot 100 = 12\%$$

The completed table is shown here.

A Class boundaries	B Frequency	C Cumulative frequency	D Cumulative percent
89.5–104.5	24	24	12
104.5–119.5	62	86	43
119.5–134.5	72	158	79
134.5–149.5	26	184	92
149.5–164.5	12	196	98
164.5–179.5	4	200	100
	<u>200</u>		

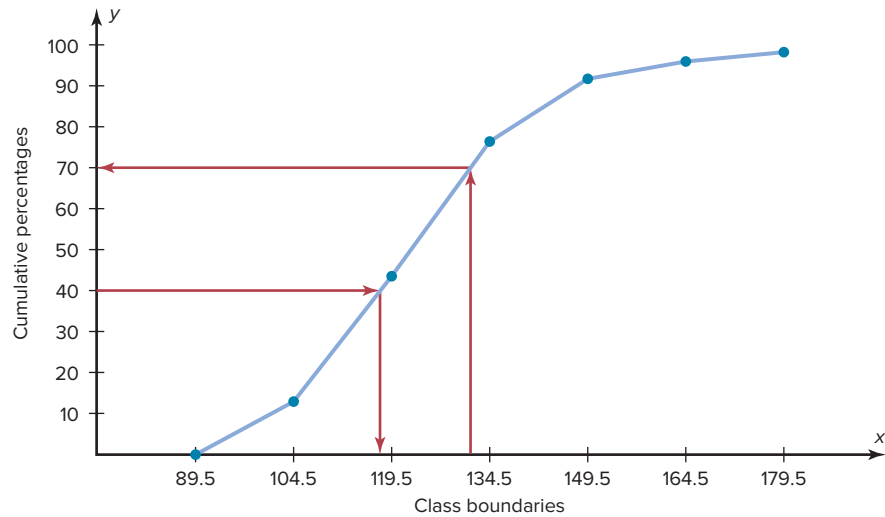
**Step 3** Graph the data, using class boundaries for the  $x$  axis and the percentages for the  $y$  axis, as shown in Figure 3–6.

Once a percentile graph has been constructed, one can find the approximate corresponding percentile ranks for given blood pressure values and find approximate blood pressure values for given percentile ranks.

For example, to find the percentile rank of a blood pressure reading of 130, find 130 on the  $x$  axis of Figure 3–6 and draw a vertical line to the graph. Then move horizontally to the value on the  $y$  axis. Note that a blood pressure of 130 corresponds to approximately the 70th percentile.

If the value that corresponds to the 40th percentile is desired, start on the  $y$  axis at 40 and draw a horizontal line to the graph. Then draw a vertical line to the  $x$  axis and read the value. In Figure 3–6, the 40th percentile corresponds to a value of approximately 118. Thus, if a person has a blood pressure of 118, he or she is at the 40th percentile.

Finding values and the corresponding percentile ranks by using a graph yields only approximate answers. Several mathematical methods exist for computing percentiles for data. These methods can be used to find the approximate percentile rank of a data value

**FIGURE 3-6**Percentile Graph for  
Example 3-29

or to find a data value corresponding to a given percentile. When the data set is large (100 or more), these methods yield better results. Examples 3-30 and 3-31 show these methods.

#### Percentile Formula

The percentile corresponding to a given value  $X$  is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

#### EXAMPLE 3-30 Traffic Violations

The number of traffic violations recorded by a police department for a 10-day period is shown. Find the percentile rank of 16.

22 19 25 24 18 15 9 12 16 20

#### SOLUTION

Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

Since there are 3 numbers below the value of 16, the solution is

$$\text{Percentile} = \frac{3 + 0.5}{10} \cdot 100 = 35\text{th percentile}$$

Hence, the value of 16 is higher than 35% of the data values.

*Note:* One assumes that a value of 16, for instance, means theoretically any value between 15.5 and 16.5.

**EXAMPLE 3–31** Traffic Violations

Using the data in Example 3–30, find the percentile rank of 24.

**SOLUTION**

There are 8 values below 24; thus,

$$\text{Percentile} = \frac{8 + 0.5}{10} \cdot 100 = 85\text{th percentile}$$

Therefore, the data value 24 is higher than 85% of the values in the data set.

The steps for finding a value corresponding to a given percentile are summarized in this Procedure Table.

**Procedure Table****Finding a Data Value Corresponding to a Given Percentile**

**Step 1** Arrange the data in order from lowest to highest.

**Step 2** Substitute into the formula

$$c = \frac{n \cdot p}{100}$$

where  $n$  = total number of values

$p$  = percentile

**Step 3A** If  $c$  is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

**Step 3B** If  $c$  is a whole number, use the value halfway between the  $c$ th and  $(c + 1)$ st values when counting up from the lowest value.

Examples 3–32 and 3–33 show a procedure for finding a value corresponding to a given percentile.

**EXAMPLE 3–32** Traffic Violations

Using the data in Example 3–30, find the value corresponding to the 65th percentile.

**SOLUTION**

**Step 1** Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

**Step 2** Compute

$$c = \frac{n \cdot p}{100}$$

where  $n$  = total number of values

$p$  = percentile

Thus,

$$c = \frac{10 \cdot 65}{100} = 6.5$$

Since  $c$  is not a whole number, round it up to the next whole number; in this case, it is  $c = 7$ .

Start at the lowest value and count over to the 7th value, which is 20. Hence, the value of 20 corresponds to the 65th percentile.



**EXAMPLE 3–33** Traffic Violations

Using the data in Example 3–30, find the data value corresponding to the 30th percentile.

**SOLUTION**

**Step 1** Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

**Step 2** Substitute in the formula.

$$c = \frac{n \cdot p}{100} \quad c = \frac{10 \cdot 30}{100} = 3$$

In this case, it is the 3rd and 4th values.

**Step 3** Since  $c$  is a whole number, use the value halfway between the  $c$  and  $c + 1$  values when counting up from the lowest. In this case, it is the third and fourth values.

9 12 15 16 18 19 20 22 24 25  
           ↑    ↑  
       3rd 4th  
       value value

The halfway value is between 15 and 16. It is 15.5. Hence, 15.5 corresponds to the 30th percentile.

**Quartiles and Deciles**

**Quartiles** divide the distribution into four equal groups, denoted by  $Q_1$ ,  $Q_2$ ,  $Q_3$ .

Note that  $Q_1$  is the same as the 25th percentile;  $Q_2$  is the same as the 50th percentile, or the median;  $Q_3$  corresponds to the 75th percentile, as shown:



Quartiles can be computed by using the formula given for computing percentiles on page 153. For  $Q_1$  use  $p = 25$ . For  $Q_2$  use  $p = 50$ . For  $Q_3$  use  $p = 75$ . However, an easier method for finding quartiles is found in this Procedure Table.

**Procedure Table****Finding Data Values Corresponding to  $Q_1$ ,  $Q_2$ , and  $Q_3$** 

- Step 1** Arrange the data in order from lowest to highest.
- Step 2** Find the median of the data values. This is the value for  $Q_2$ .
- Step 3** Find the median of the data values that fall below  $Q_2$ . This is the value for  $Q_1$ .
- Step 4** Find the median of the data values that fall above  $Q_2$ . This is the value for  $Q_3$ .

Example 3–34 shows how to find the values of  $Q_1$ ,  $Q_2$ , and  $Q_3$ .

**EXAMPLE 3–34** Traffic Violations

Using the data in Exercise 3–30, find  $Q_1$ ,  $Q_2$ , and  $Q_3$ .

**SOLUTION**

**Step 1** Arrange the data in order from lowest to highest.

9 12 15 16 18 19 20 22 24 25

**Step 2** Find the median  $Q_2$ .

9 12 15 16 18 19 20 22 24 25

$$\begin{array}{c} \uparrow \\ \text{MD} = \frac{18 + 19}{2} = 18.5 \end{array}$$

**Step 3** Find the median of the data values below 18.5.

9 12 15 16 18

$$\begin{array}{c} \uparrow \\ Q_1 = 15 \end{array}$$

**Step 4** Find the median of the data values greater than 18.5.

19 20 22 24 25

$$\begin{array}{c} \uparrow \\ Q_3 = 22 \end{array}$$

Hence,  $Q_1 = 15$ ,  $Q_2 = 18.5$ , and  $Q_3 = 22$ .

In addition to dividing the data set into four groups, quartiles can be used as a rough measure of variability. This measure of variability which uses quartiles is called the *interquartile range* and is the range of the middle 50% of the data values.

The **interquartile range (IQR)** is the difference between the third and first quartiles.

$$\text{IQR} = Q_3 - Q_1$$

**EXAMPLE 3–35** Traffic Violations

Find the interquartile range of the data set in Example 3–30.

**SOLUTION**

Find  $Q_1$  and  $Q_3$ . This was done in Example 3–34. Now  $Q_1 = 15$  and  $Q_3 = 22$ . Next subtract  $Q_1$  from  $Q_3$ .

$$\text{IQR} = Q_3 - Q_1 = 22 - 15 = 7$$

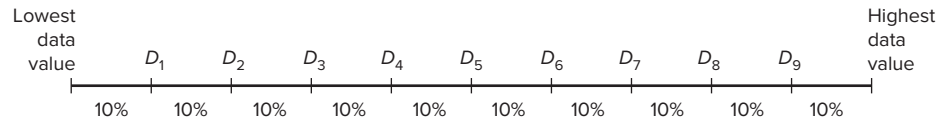
The interquartile range is equal to 7.

Like the standard deviation, the more variable the data set is, the larger the value of the interquartile range will be.

**Deciles** divide the distribution into 10 groups, as shown. They are denoted by  $D_1$ ,  $D_2$ , etc.

### Unusual Stat

Of the alcoholic beverages consumed in the United States, 85% is beer.



Note that  $D_1$  corresponds to  $P_{10}$ ;  $D_2$  corresponds to  $P_{20}$ ; etc. Deciles can be found by using the formulas given for percentiles. Taken altogether then, these are the relationships among percentiles, deciles, and quartiles.

Deciles are denoted by  $D_1, D_2, D_3, \dots, D_9$ , and they correspond to  $P_{10}, P_{20}, P_{30}, \dots, P_{90}$ .

Quartiles are denoted by  $Q_1, Q_2, Q_3$  and they correspond to  $P_{25}, P_{50}, P_{75}$ .

The median is the same as  $P_{50}$  or  $Q_2$  or  $D_5$ .

The position measures are summarized in Table 3–4.

**TABLE 3–4 Summary of Position Measures**

Measure	Definition	Symbol(s)
Standard score or z score	Number of standard deviations that a data value is above or below the mean	$z$
Percentile	Position in hundredths that a data value holds in the distribution	$P_n$
Decile	Position in tenths that a data value holds in the distribution	$D_n$
Quartile	Position in fourths that a data value holds in the distribution	$Q_n$

### Outliers

A data set should be checked for extremely high or extremely low values. These values are called *outliers*.

An **outlier** is an extremely high or an extremely low data value when compared with the rest of the data values.

An outlier can strongly affect the mean and standard deviation of a variable. For example, suppose a researcher mistakenly recorded an extremely high data value. This value would then make the mean and standard deviation of the variable much larger than they really were.

Since these measures (mean and standard deviation) are affected by outliers, they are called *nonresistant statistics*. The median and interquartile range are less affected by outliers, so they are called *resistant statistics*. Sometimes when a distribution is skewed or contains outliers, the median and interquartile range can be used to more accurately describe the data than the mean and standard deviation. Outliers can have an effect on other statistics as well.

There are several ways to check a data set for outliers. One method is shown in this Procedure Table.

### Procedure Table

#### Procedure for Identifying Outliers

- |               |   |
|---------------|---|
| <b>Step 1</b> | Arrange the data in order from lowest to highest and find $Q_1$ and $Q_3$ .                                   |
| <b>Step 2</b> | Find the interquartile range: $IQR = Q_3 - Q_1$ .   |
| <b>Step 3</b> | Multiply the IQR by 1.5.  |
| <b>Step 4</b> | Subtract the value obtained in step 3 from $Q_1$ and add the value obtained in step 3 to $Q_3$ .              |
| <b>Step 5</b> | Check the data set for any data value that is smaller than $Q_1 - 1.5(IQR)$ or larger than $Q_3 + 1.5(IQR)$ . |

This procedure is shown in Example 3–36.

### EXAMPLE 3–36 Outliers

Check the following data set for outliers.

5, 6, 12, 13, 15, 18, 22, 50

#### SOLUTION

The data value 50 is extremely suspect. These are the steps in checking for an outlier.

**Step 1** Find  $Q_1$  and  $Q_3$ .  $Q_1 = \frac{(6 + 12)}{2} = 9$ ;  $Q_3 = \frac{(18 + 22)}{2} = 20$ .

**Step 2** Find the interquartile range (IQR), which is  $Q_3 - Q_1$ .

$$IQR = Q_3 - Q_1 = 20 - 9 = 11$$

**Step 3** Multiply this value by 1.5.

$$1.5(11) = 16.5$$

**Step 4** Subtract the value obtained in step 3 from  $Q_1$ , and add the value obtained in step 3 to  $Q_3$ .

$$9 - 16.5 = -7.5 \quad \text{and} \quad 20 + 16.5 = 36.5$$

**Step 5** Check the data set for any data values that fall outside the interval from  $-7.5$  to  $36.5$ . The value 50 is outside this interval; hence, it can be considered an outlier.

There are several reasons why outliers may occur. First, the data value may have resulted from a measurement or observational error. Perhaps the researcher measured the variable incorrectly. Second, the data value may have resulted from a recording error. That is, it may have been written or typed incorrectly. Third, the data value may have been obtained from a subject that is not in the defined population. For example, suppose test scores were obtained from a seventh-grade class, but a student in that class was actually in the sixth grade and had special permission to attend the class. This student might have scored extremely low on that particular exam on that day. Fourth, the data value might be a legitimate value that occurred by chance (although the probability is extremely small).

There are no hard-and-fast rules on what to do with outliers, nor is there complete agreement among statisticians on ways to identify them. Obviously, if they occurred as a result of an error, an attempt should be made to correct the error or else the data value should be omitted entirely. When they occur naturally by chance, the statistician must make a decision about whether to include them in the data set.

When a distribution is normal or bell-shaped, data values that are beyond 3 standard deviations of the mean can be considered suspected outliers.

## Applying the Concepts 3–3

### Determining Dosages

In an attempt to determine necessary dosages of a new drug (HDL) used to control sepsis, assume you administer varying amounts of HDL to 40 mice. You create four groups and label them *low dosage*, *moderate dosage*, *large dosage*, and *very large dosage*. The dosages also vary within each group. After the mice are injected with the HDL and the sepsis bacteria, the time until the onset of sepsis is recorded. Your job as a statistician is to effectively communicate the results of the study.

1. Which measures of position could be used to help describe the data results?
2. If 40% of the mice in the top quartile survived after the injection, how many mice would that be?
3. What information can be given from using percentiles?
4. What information can be given from using quartiles?
5. What information can be given from using standard scores?

See page 184 for the answers.

## Exercises 3–3

1. What is a  $z$  score?
2. Define *percentile rank*.
3. What is the difference between a percentage and a percentile?
4. Define *quartile*.
5. What is the relationship between quartiles and percentiles?
6. What is a decile?
7. How are deciles related to percentiles?
8. To which percentile, quartile, and decile does the median correspond?
9. **Vacation Days** If the average number of vacation days for a selection of various countries has a mean of 29.4 days and a standard deviation of 8.6 days, find the  $z$  scores for the average number of vacation days in each of these countries.  

Canada	26 days
Italy	42 days
United States	13 days

Source: www.infoplease.com
10. **Age of Senators** The average age of Senators in the 114th congress was 61.7 years. If the standard deviation was 10.6, find the  $z$  scores of a senator who is 48 years old and one who is 66 years old.
11. **Marriage Age for Females** The mean age at which females marry is 24.6. The standard deviation is 3.2 years. Find the corresponding  $z$  score for each.  

a. 27	d. 18
b. 22	e. 26
c. 31	
12. **Teacher's Salary** The average teacher's salary in a particular state is \$54,166. If the standard deviation is \$10,200, find the salaries corresponding to the following  $z$  scores.  

a. 2	d. 2.5
b. -1	e. -1.6
c. 0	
13. **Test Scores** Which is a better relative position, a score of 83 on a geography test that has a mean of 72 and a standard deviation of 6, or a score of 61 on an accounting test that has a mean of 55 and a standard deviation of 3.5?
14. **College and University Debt** A student graduated from a 4-year college with an outstanding loan of \$9650 where the average debt is \$8455 with a standard deviation of \$1865. Another student graduated from a university with an outstanding loan of \$12,360 where the average of the outstanding loans was \$10,326 with a standard deviation of \$2143. Which student had a higher debt in relationship to his or her peers?

- 15. Annual Miles Driven** The average miles driven annually per licensed driver in the United States is approximately 14,090 miles. If we assume a fairly mound-shaped distribution with a standard deviation of approximately 3500 miles, find the following:

- $z$  score for 16,000 miles
- $z$  score for 10,000 miles
- Number of miles corresponding to  $z$  scores of 1.6,  $-0.5$ , and 0.

Source: World Almanac 2012.

- 16.** Which score indicates the highest relative position?

- A score of 3.2 on a test with  $\bar{X} = 4.6$  and  $s = 1.5$
- A score of 630 on a test with  $\bar{X} = 800$  and  $s = 200$
- A score of 43 on a test with  $\bar{X} = 50$  and  $s = 5$

- 17.** The data show the population (in thousands) for a recent year of a sample of cities in South Carolina.

29	26	15	13	17	58
14	25	37	19	40	67
23	10	97	12	129	
27	20	18	120	35	
66	21	11	43	22	

Source: U.S. Census Bureau.

Find the data value that corresponds to each percentile.

- 40th percentile
- 75th percentile
- 90th percentile
- 30th percentile

Using the same data, find the percentile corresponding to the given data value.

- 27
- 40
- 58
- 67

- 18. College Room and Board Costs** Room and board costs for selected schools are summarized in this distribution. Find the approximate cost of room and board corresponding to each of the following percentiles by constructing a percentile graph.

Costs (in dollars)	Frequency
3000.5–4000.5	5
4000.5–5000.5	6
5000.5–6000.5	18
6000.5–7000.5	24
7000.5–8000.5	19
8000.5–9000.5	8
9000.5–10,000.5	5

- 30th percentile
- 50th percentile
- 75th percentile
- 90th percentile

Source: World Almanac.

Using the same data, find the approximate percentile rank of each of the following costs.

- 5500
- 6500
- 7200
- 8300

- 19. Achievement Test Scores** The data shown represent the scores on a national achievement test for a group of 10th-grade students. Find the approximate percentile ranks of these scores by constructing a percentile graph.

- 220
- 245
- 276
- 280
- 300

Score	Frequency
196.5–217.5	5
217.5–238.5	17
238.5–259.5	22
259.5–280.5	48
280.5–301.5	22
301.5–322.5	6

For the same data, find the approximate scores that correspond to these percentiles.

- 15th
- 29th
- 43rd
- 65th
- 80th

- 20. Airplane Speeds** The airborne speeds in miles per hour of 21 planes are shown. Find the approximate values that correspond to the given percentiles by constructing a percentile graph.

Class	Frequency
366–386	4
387–407	2
408–428	3
429–449	2
450–470	1
471–491	2
492–512	3
513–533	4
	21

Source: The World Almanac and Book of Facts.

- 9th
- 20th
- 45th
- 60th
- 75th

Using the same data, find the approximate percentile ranks of the following speeds in miles per hour (mph).

- f. 380 mph                      i. 505 mph  
g. 425 mph                      j. 525 mph  
h. 455 mph

- 21. Average Weekly Earnings** The average weekly earnings in dollars for various industries are listed below. Find the percentile rank of each value.

804 736 659 489 777 623 597 524 228

For the same data, what value corresponds to the 40th percentile?

Source: *New York Times Almanac*.

- 22. Test Scores** Find the percentile rank for each test score in the data set.

12, 28, 35, 42, 47, 49, 50

What value corresponds to the 60th percentile?

- 23. Hurricane Damage** Find the percentile rank for each value in the data set. The data represent the values in billions of dollars of the damage of 10 hurricanes.

1.1, 1.7, 1.9, 2.1, 2.2, 2.5, 3.3, 6.2, 6.8, 20.3

What value corresponds to the 40th percentile?

Source: *Insurance Services Office*.

- 24. Test Scores** Find the percentile rank for each test score in the data set.

5, 12, 15, 16, 20, 21

What test score corresponds to the 33rd percentile?

- 25. Taxes** The data for a recent year show the taxes (in millions of dollars) received from a random sample of 10 states. Find the first and third quartiles and the IQR.

13 15 32 36 11 24 6 25 11 71

Source: U.S. Census Bureau.

- 26. Medical Marijuana 2015 Sales Tax:** The data show the amount of sales tax paid in Denver County, Colorado. Find the first and third quartiles for the data.

Month	Sales Tax	Month	Sales Tax
Jan	363,061	July	518,868
Feb	358,208	August	554,013
March	418,500	September	506,809
April	266,771	October	341,421
May	399,814	November	349,026
June	453,698	December	532,545

Source: Colorado Department of Revenue

- 27. Gold Reserves** The data show the gold reserves for a recent year for 11 world countries. Find the first and third quartiles and the IQR. The data are in millions of troy ounces.

33.9 78.3 108.9 17.9 78.8 24.6 19.7 33.3  
33.4 10.0 261.5

Source: International Monetary Fund.

- 28. Police Calls in Schools** The number of incidents in which police were needed for a sample of 9 schools in Allegheny County is 7, 37, 3, 8, 48, 11, 6, 0, 10. Find the first and third quartiles for the data.

- 29.** Check each data set for outliers.

- a. 46, 28, 32, 21, 25, 29, 34, 19  
b. 82, 100, 97, 93, 89, 90, 65, 94, 101  
c. 527, 1007, 489, 371, 175

- 30.** Check each data set for outliers.

- a. 88, 72, 97, 84, 86, 85, 100  
b. 145, 119, 122, 118, 125, 116  
c. 14, 16, 27, 18, 13, 19, 36, 15, 20

## Extending the Concepts

- 31.** Another measure of the average is called the *midquartile*; it is the numerical value halfway between  $Q_1$  and  $Q_3$ , and the formula is

$$\text{Midquartile} = \frac{Q_1 + Q_3}{2}$$

Using this formula and other formulas, find  $Q_1$ ,  $Q_2$ ,  $Q_3$ , the midquartile, and the interquartile range for each data set.

- a. 5, 12, 16, 25, 32, 38  
b. 53, 62, 78, 94, 96, 99, 103

- 32.** An employment evaluation exam has a variance of 250. Two particular exams with raw scores of 142 and 165 have  $z$  scores of  $-0.5$  and  $0.955$ , respectively. Find the mean of the distribution.

- 33.** A particular standardized test has scores that have a mound-shaped distribution with mean equal to 125 and standard deviation equal to 18. Tom had a raw score of 158, Dick scored at the 98th percentile, and Harry had a  $z$  score of 2.00. Arrange these three students in order of their scores from lowest to highest. Explain your reasoning.



## Technology

TI-84 Plus  
Step by Step

## Step by Step

## Calculating Descriptive Statistics

To calculate various descriptive statistics:

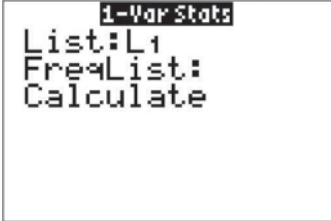
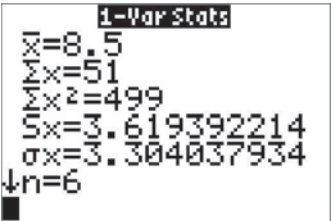
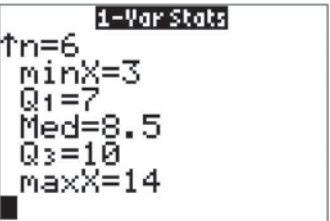
1. Enter data into L1.
2. Press **STAT** to get the menu.
3. Press **→** to move cursor to **CALC**; then press **1** for 1-Var Stats.
4. Under List, press **2nd [L1]**, then **ENTER**.
5. Leave FreqList blank, then press **ENTER**.
6. While highlighting Calculate, press **ENTER**.

The calculator will display

$\bar{x}$	sample mean
$\sum x$	sum of the data values
$\sum x^2$	sum of the squares of the data values
$S_x$	sample standard deviation
$\sigma_x$	population standard deviation
$n$	number of data values
minX	smallest data value
$Q_1$	lower quartile
Med	median
$Q_3$	upper quartile
maxX	largest data value

## Example TI3-1

Find the various descriptive statistics for the teacher strikes data from Example 3-20: 9, 10, 14, 7, 8, 3

Input	Output	Output
		

Following the steps just shown, we obtain these results, as shown on the screen:

- The mean is 8.5.
- The sum of  $x$  is 51.
- The sum of  $x^2$  is 499.
- The sample standard deviation  $S_x$  is 3.619392214.
- The population standard deviation  $\sigma_x$  is 3.304037934.
- The sample size  $n$  is 6.
- The smallest data value is 3.
- $Q_1$  is 7.
- The median is 8.5.
- $Q_3$  is 10.
- The largest data value is 14.

To calculate the mean and standard deviation from grouped data:

1. Enter the midpoints into **L1**.
2. Enter the frequencies into **L2**.
3. Press **STAT** to get the menu.
4. Use the arrow keys to move the cursor to **CALC**; then press 1 for 1-Var Stats.
5. Under List, press **2nd [L1]**, then **ENTER**.
6. Under FreqList press **2nd [L2]**, then **ENTER**.
7. While highlighting Calculate, press **ENTER**.

#### Example TI3–2

Calculate the mean and standard deviation for the data given in Examples 3–3 and 3–22.

Class	Frequency	Midpoint
5.5–10.5	1	8
10.5–15.5	2	13
15.5–20.5	3	18
20.5–25.5	5	23
25.5–30.5	4	28
30.5–35.5	3	33
35.5–40.5	2	38

Input

L1	L2	L3	1
8	1	-----	
13	2		
18	3		
23	5		
28	4		
33	3		
38	2		
L1(1)=8			

Input

1-Var Stats
List:L1
FreqList:L2
Calculate

Output

1-Var Stats
$\bar{x}=24.5$
$\Sigma x=490$
$\Sigma x^2=13310$
$Sx=8.287593772$
$\sigma x=8.077747211$
$n=20$

The sample mean is 24.5, and the sample standard deviation is 8.287593772.

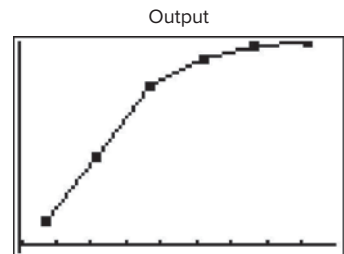
To graph a percentile graph, follow the procedure for an ogive (Section 2–2), but use the cumulative percent in L2, 100 for Ymax, and the data from Example 3–29.

Input

L1	L2	L3	2
97	12	-----	
112	43		
127	79		
142	92		
157	98		
172	100		
L2(7) =			

Input

Plot1 Plot2 Plot3
On Off
Type: L1 L2 L3
Xlist:L1
Ylist:L2
Mark:



## EXCEL

### Step by Step


## Measures of Position

### Example XL3–3

Find the  $z$  scores for each value of the data from Example 3–36.

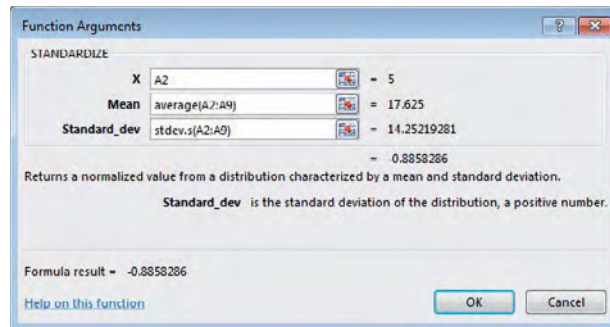
5   6   12   13   15   18   22   50

1. On an Excel worksheet enter the data in cells A2–A9. Enter a label for the variable in cell A1.
2. Label cell B1 as  $z$  score.

3. Select cell B2.
4. Select the Formulas tab from the toolbar and Insert Function .
5. Select the Statistical category for statistical functions and scroll in the function list to STANDARDIZE and click [OK].

In the STANDARDIZE dialog box:

6. Type A2 for the X value.
7. Type average(A2:A9) for the mean.
8. Type stdev.s(A2:A9) for the Standard\_dev. Then click [OK].
9. Repeat the procedure above for each data value in column A.



#### Example XL3–4


Excel has two built-in functions to find the Percentile Rank corresponding to a value in a set of data.

PERCENTRANK.INC calculates the Percentile Rank corresponding to a data value in the range 0 to 1 inclusively.

PERCENTRANK.EXC calculates the Percentile Rank corresponding to a data value in the range 0 to 1 exclusively.

We will compute Percentile Ranks for the data from Example 3–36, using both PERCENTRANK.INC and PERCENTRANK.EXC to demonstrate the difference between the two functions.

5    6    12    13    15    18    22    50

1. On an Excel worksheet enter the data in cells A2–A9. Enter the label **Data** in cell A1.
2. Label cell B1 as **Percent Rank INC** and cell C1 as **Percent Rank EXC**.
3. Select cell B2.
4. Select the Formulas tab from the toolbar and Insert Function .
5. Select the Statistical category for statistical functions and scroll in the function list to PERCENTRANK.INC (PERCENTRANK.EXC) and click [OK].

In the PERCENTRANK.INC (PERCENTRANK.EXC) dialog boxes:

6. Type A2:A9 for the Array.
7. Type A2 for X, then click [OK]. You can leave the Significance box blank unless you want to change the number of significant digits of the output (the default is 3 significant digits).
8. Repeat the procedure above for each data value in the set.

The function results for both PERCENTRANK.INC and PERCENTRANK.EXC are shown below.

*Note:* Both functions return the Percentile Ranks as a number between 0 and 1. You may convert these to numbers between 0 and 100 by multiplying each function value by 100.

The screenshot shows an Excel spreadsheet with the following data:

	A	B	C	D	E	F	G	H	I	J	K
1	Data	Percent Rank INC	Percent Rank EXC								
2	5	0	0.111								
3	6	0.142	0.222								
4	12	0.285	0.333								
5	13	0.428	0.444								
6	15	0.571	0.555								
7	18	0.714	0.666								
8	22	0.857	0.777								
9	50	1	0.888								

## Descriptive Statistics in Excel

### Example XL3–5

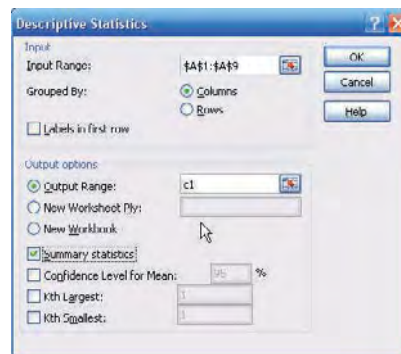
Excel Analysis Tool-Pak Add-in Data Analysis includes an item called Descriptive Statistics that reports many useful measures for a set of data.

1. Enter the data set shown in cells A1 to A9 of a new worksheet.

12 17 15 16 16 14 18 13 10

See the Excel Step by Step in Chapter 1 for the instructions on loading the Analysis Tool-Pak Add-in.

2. Select the Data tab on the toolbar and select Data Analysis.
3. In the Analysis Tools dialog box, scroll to Descriptive Statistics, then click [OK].
4. Type A1:A9 in the Input Range box and check the Grouped by Columns option.
5. Select the Output Range option and type in cell C1.
6. Check the Summary statistics option and click [OK].



Below is the summary output for this data set.

Column1	
Mean	14.55555556
Standard Error	0.85165054
Median	15
Mode	16
Standard Deviation	2.554951619
Sample Variance	6.527777778
Kurtosis	-0.3943866
Skewness	-0.51631073
Range	8
Minimum	10
Maximum	18
Sum	131
Count	9

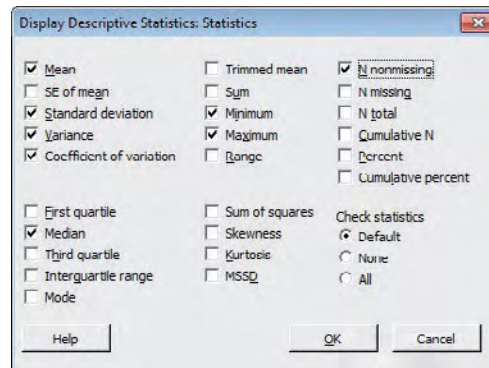
## MINITAB

### Step by Step

### Calculate Descriptive Statistics from Data

#### Example MT3–1

1. Enter the data from Example 3–20 on teacher strikes into C1 of MINITAB. Name the column **Strikes**.
2. Select **Stat>Basic Statistics>Display Descriptive Statistics**.
3. The cursor will be blinking in the Variables text box. Double-click C1 Strikes.
4. Click [Statistics] to view the statistics that can be calculated with this command.
  - a) Check the boxes for Mean, Standard deviation, Variance, Coefficient of variation, Median, Minimum, Maximum, and N nonmissing.



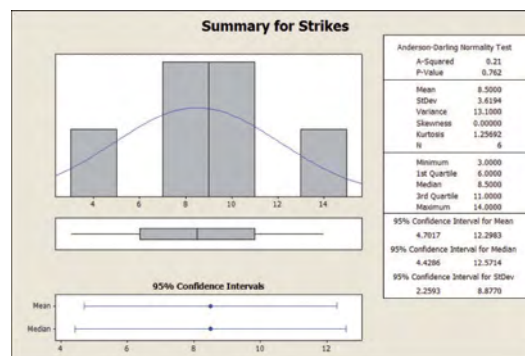
- b) Remove the checks from other options.
5. Click [OK] twice. The results will be displayed in the session window as shown.

#### Descriptive Statistics: Strikes

Variable	N	Mean	StDev	Variance	CoefVar	Minimum	Median	Maximum
Strikes	6	8.50	3.62	13.10	42.58	3.00	8.50	14.00

Session window results are in text format. A high-resolution graphical window displays the descriptive statistics, a histogram, and a boxplot.

6. Select **Stat>Basic Statistics>Graphical Summary**.
7. Double-click C1 Strikes.
8. Click [OK].



The graphical summary will be displayed in a separate window as shown.

### Calculate Descriptive Statistics from a Frequency Distribution

Multiple menu selections must be used to calculate the statistics from a table. We will use data given in Example 3–22 on miles run per week.

### Enter Midpoints and Frequencies

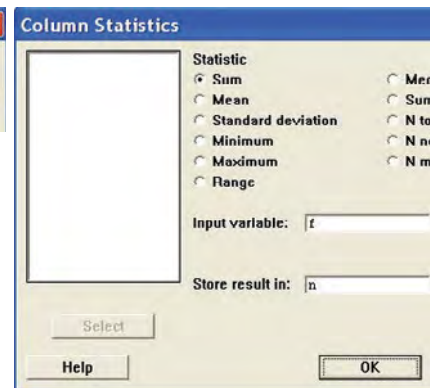
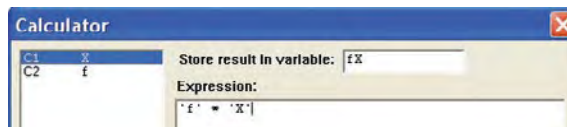
1. Select **File>New>New Worksheet** to open an empty worksheet.
2. To enter the midpoints into C1, select **Calc>Make Patterned Data>Simple Set of Numbers**.
  - a) Type **X** to name the column.
  - b) Type in **8** for the First value, **38** for the Last value, and **5** for Steps.
  - c) Click [OK].
3. Enter the frequencies in C2. Name the column **f**.

### Calculate Columns for $f \cdot X$ and $f \cdot X^2$

4. Select **Calc>Calculator**.
  - a) Type in **fX** for the variable and **f\*X** in the Expression dialog box. Click [OK].
  - b) Select **Edit>Edit Last Dialog** and type in **fX2** for 'Store Result in Variable' and **f\*X\*\*2** for the expression.
  - c) Click [OK]. There are now four columns in the worksheet.

### Calculate the Column Sums

5. Select **Calc>Column Statistics**.  
This command stores results in constants, not columns.  
Click [OK] after each step.
  - a) Click the option for **Sum**; then select C2 **f** for the Input column, and type **n** for Store result in.
  - b) Select **Edit>Edit Last Dialog**; then select C3 **fX** for the column and type **sumX** for storage.

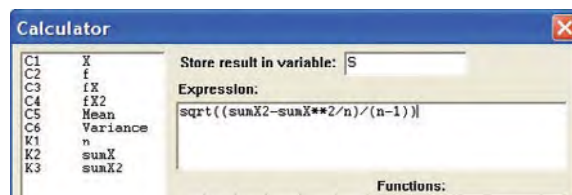


- c) Edit the last dialog box again. This time select C4 **fX2** for the column, then type **sumX2** for storage.

To verify the results, navigate to the Project Manager window, then the constants folder of the worksheet. The sums are 20, 490, and 13,310.

### Calculate the Mean, Variance, and Standard Deviation

6. Select **Calc>Calculator**.
  - a) Type **Mean** for the variable, then click in the box for the Expression and type **sumX/n**. Click [OK]. If you double-click the constants instead of typing them, single quotes will surround the names. The quotes are not required unless the column name has spaces.
  - b) Click the **EditLast Dialog** icon and type **Variance** for the variable.
  - c) In the expression box type in 
$$\frac{(\text{sumX2} - \text{sumX}^2/n)}{(n-1)}$$



- d) Edit the last dialog box and type **S** for the variable. In the expression box, drag the mouse over the previous expression to highlight it.
- e) Click the button in the keypad for parentheses. Type **SQRT** at the beginning of the line, upper- or lowercase will work. The expression should be  $\text{SQRT}((\text{sumX2} - \text{sumX}^2/n)/(n-1))$ .
- f) Click [OK].

### Display Results

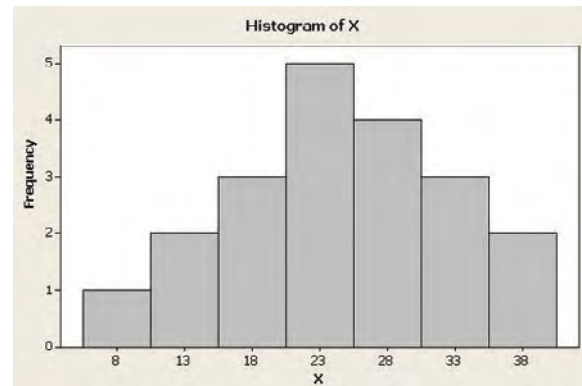
- g) Select **Data>Display Data**, then highlight all columns and constants in the list.
- h) Click [Select] then [OK].

The session window will display all our work! Create the histogram with instructions from Chapter 2.

#### Data Display

n 20.0000  
sumX 490.000  
sumX2 13310.0

Row	X	f	fX	fX2	Mean	Variance	S
1	8	1	8	64	24.5	68.6842	8.28759
2	13	2	26	338			
3	18	3	54	972			
4	23	5	115	2645			
5	28	4	112	3136			
6	33	3	99	3267			
7	38	2	76	2888			



## 3-4 Exploratory Data Analysis

### OBJECTIVE 4

Use the techniques of exploratory data analysis, including boxplots and five-number summaries, to discover various aspects of data.

In traditional statistics, data are organized by using a frequency distribution. From this distribution various graphs such as the histogram, frequency polygon, and ogive can be constructed to determine the shape or nature of the distribution. In addition, various statistics such as the mean and standard deviation can be computed to summarize the data.

The purpose of traditional analysis is to confirm various conjectures about the nature of the data. For example, from a carefully designed study, a researcher might want to know if the proportion of Americans who are exercising today has increased from 10 years ago. This study would contain various assumptions about the population, various definitions such as the definition of exercise, and so on.

In **exploratory data analysis (EDA)**, data can be organized using a *stem and leaf plot*. (See Chapter 2.) The measure of central tendency used in EDA is the *median*. The measure of variation used in EDA is the *interquartile range*  $Q_3 - Q_1$ . In EDA the data are represented graphically using a *boxplot* (sometimes called a box and whisker plot). The purpose of exploratory data analysis is to examine data to find out what information can be discovered about the data, such as the center and the spread. Exploratory data analysis was developed by John Tukey and presented in his book *Exploratory Data Analysis* (Addison-Wesley, 1977).

### The Five-Number Summary and Boxplots

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum)
2.  $Q_1$
3. The median



- 4.  $Q_3$
- 5. The highest value of the data set (i.e., maximum)

These values are called a **five-number summary** of the data set.

A **boxplot** is a graph of a data set obtained by drawing a horizontal line from the minimum data value to  $Q_1$ , drawing a horizontal line from  $Q_3$  to the maximum data value, and drawing a box whose vertical sides pass through  $Q_1$  and  $Q_3$  with a vertical line inside the box passing through the median or  $Q_2$ .

Procedure Table

Constructing a Boxplot

Step 1

Find the five-number summary for the data.

Step 2

Draw a horizontal axis and place the scale on the axis. The scale should start on or below the minimum data value and end on or above the maximum data value.

Step 3

Locate the lowest data value,  $Q_1$ , the median,  $Q_3$ , and the highest data value; then draw a box whose vertical sides go through  $Q_1$  and  $Q_3$ . Draw a vertical line through the median. Finally, draw a line from the minimum data value to the left side of the box, and draw a line from the maximum data value to the right side of the box.

EXAMPLE 3–37 Number of Meteorites Found

The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.  
*Source:* Natural History Museum.

SOLUTION

- Step 1
- Find the five-number summary for the data.  
Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Find the median.

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

↑  
Median

$Median = \frac{78 + 89}{2} = 83.5$

Find  $Q_1$ .

30, 39, 47, 48, 78

↑  
 $Q_1$

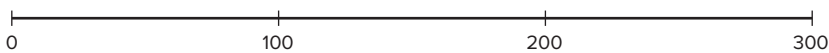
Find  $Q_3$ .

89, 138, 164, 215, 296

↑  
 $Q_3$

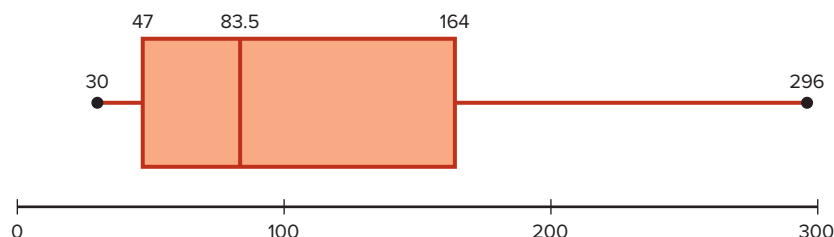
The minimum data value is 30, and the maximum data value is 296.

- Step 2
- Draw a horizontal axis and the scale.



**Step 3** Draw the box above the scale using  $Q_1$  and  $Q_3$ . Draw a vertical line through the median, and draw lines from the lowest data value to the box and from the highest data value to the box. See Figure 3–7.

**FIGURE 3–7** Boxplot for Example 3–37



#### Information Obtained from a Boxplot

1. a. If the median is near the center of the box, the distribution is approximately symmetric.  
 b. If the median falls to the left of the center of the box, the distribution is positively skewed.  
 c. If the median falls to the right of the center, the distribution is negatively skewed.
2. a. If the lines are about the same length, the distribution is approximately symmetric.  
 b. If the right line is larger than the left line, the distribution is positively skewed.  
 c. If the left line is larger than the right line, the distribution is negatively skewed.

The boxplot in Figure 3–7 indicates that the distribution is slightly positively skewed.

If the boxplots for two or more data sets are graphed on the same axis, the distributions can be compared. To compare the averages, use the location of the medians. To compare the variability, use the interquartile range, i.e., the length of the boxes. Example 3–38 shows this procedure.

#### EXAMPLE 3–38 Speeds of Roller Coasters

The data shown are the speeds in miles per hour of a sample of wooden roller coasters and a sample of steel roller coasters. Compare the distributions by using boxplots.

Wood				Steel			
50	56	60	48	55	70	48	28
35	67	72	68	100	106	102	120

Source: UltimateRollerCoaster.com

#### SOLUTION

**Step 1** For the wooden coasters,

$$\begin{array}{cccccccc}
 35 & 48 & 50 & 56 & 60 & 67 & 68 & 72 \\
 & & \uparrow & & \uparrow & & \uparrow & \\
 & & Q_1 & & MD & & Q_3 & 
 \end{array}$$

$$Q_1 = \frac{48 + 50}{2} = 49 \qquad MD = \frac{56 + 60}{2} = 58 \qquad Q_3 = \frac{67 + 68}{2} = 67.5$$

**Step 2** For the steel coasters,

28 48 55 70 100 102 106 120

↑        ↑        ↑

$Q_1$     MD     $Q_3$

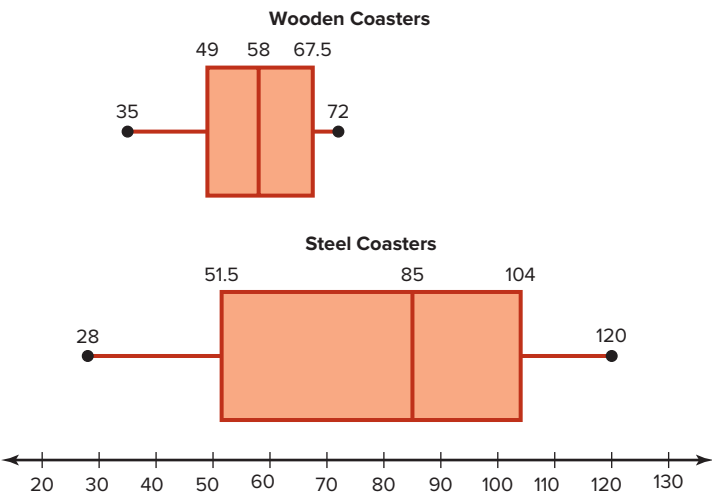
$Q_1 = \frac{48 + 55}{2} = 51.5$

$MD = \frac{70 + 100}{2} = 85$

$Q_3 = \frac{102 + 106}{2} = 104$

**Step 3** Draw the boxplots. See Figure 3–8.

**FIGURE 3–8** Boxplots for Example 3–38



The boxplots show that the median of the speeds of the steel coasters is much higher than the median speeds of the wooden coasters. The interquartile range (spread) of the steel coasters is much larger than that of the wooden coasters. Finally, the range of the speeds of the steel coasters is larger than that of the wooden coasters.

A *modified boxplot* can be drawn and used to check for outliers. See Exercise 19 in Extending the Concepts in this section.

In exploratory data analysis, *hinges* are used instead of quartiles to construct boxplots. When the data set consists of an even number of values, hinges are the same as quartiles. Hinges for a data set with an odd number of values differ somewhat from quartiles. However, since most calculators and computer programs use quartiles, they will be used in this textbook.

Table 3–5 shows the correspondence between the traditional and the exploratory data analysis approach.

TABLE 3–5 Traditional versus EDA Techniques	
Traditional	Exploratory data analysis
Frequency distribution	Stem and leaf plot
Histogram	Boxplot
Mean	Median
Standard deviation	Interquartile range

## Applying the Concepts 3–4

### The Noisy Workplace

Assume you work for OSHA (Occupational Safety and Health Administration) and have complaints about noise levels from some of the workers at a state power plant. You charge the power plant with taking decibel readings at six different areas of the plant at different times of the day and week. The results of the data collection are listed. Use boxplots to initially explore the data and make recommendations about which plant areas workers must be provided with protective ear wear. The safe hearing level is approximately 120 decibels.

Area 1	Area 2	Area 3	Area 4	Area 5	Area 6
30	64	100	25	59	67
12	99	59	15	63	80
35	87	78	30	81	99
65	59	97	20	110	49
24	23	84	61	65	67
59	16	64	56	112	56
68	94	53	34	132	80
57	78	59	22	145	125
100	57	89	24	163	100
61	32	88	21	120	93
32	52	94	32	84	56
45	78	66	52	99	45
92	59	57	14	105	80
56	55	62	10	68	34
44	55	64	33	75	21

See page 184 for the answers.

## Exercises 3–4

For Exercises 1–6, identify the five-number summary and find the interquartile range.

1. 8, 12, 32, 6, 27, 19, 54

2. 19, 16, 48, 22, 7

3. 362, 589, 437, 316, 192, 188

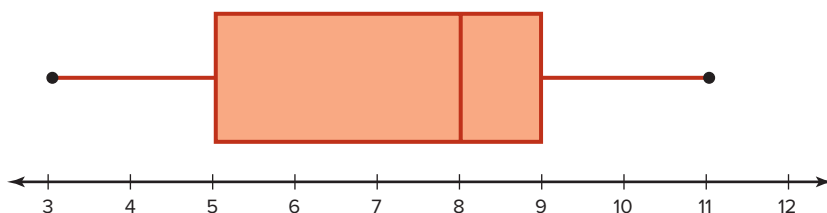
4. 147, 243, 156, 632, 543, 303

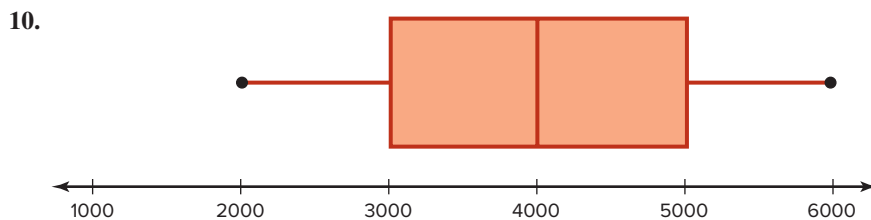
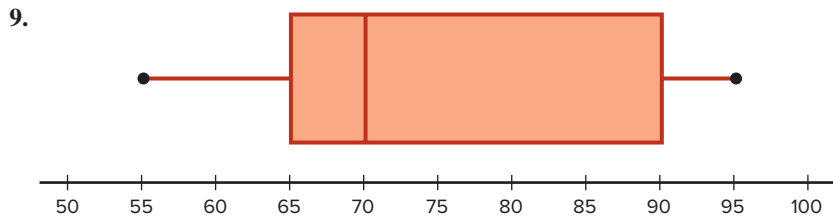
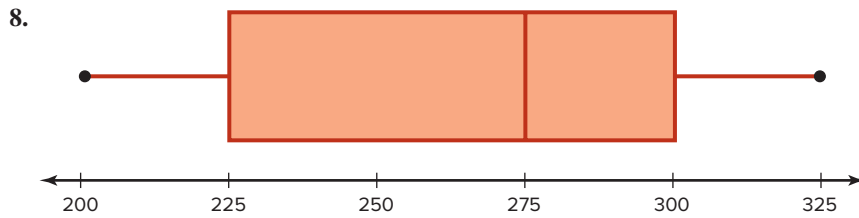
5. 14.6, 19.8, 16.3, 15.5, 18.2

6. 9.7, 4.6, 2.2, 3.7, 6.2, 9.4, 3.8

For Exercises 7–10, use each boxplot to identify the maximum value, minimum value, median, first quartile, third quartile, and interquartile range.

7.





- 11. School Graduation Rates** The data show a sample of states' percentage of public high school graduation rates for a recent year. Construct a boxplot for the data, and comment on the shape of the distribution.

79 82 77 84 80 89 60 79 91 93 88

Source: U.S. Department of Education

- 12. Innings Pitched** Construct a boxplot for the following data which represent the number of innings pitched by the ERA leaders for the past few years. Comment on the shape of the distribution.

239 266 245 236 241 246 240  
249 251 238 228 248 232

Source: Baseball-Reference.com

- 13. Population of Colonies** The data show the population (in thousands) of the U.S. Colonies in 1700 (Vermont was not a colony until 1791). Construct a boxplot and decide if the distribution is symmetric.

26.0 2.5 29.6 55.9 5.0 14.0 19.1 10.7  
18.0 5.9 5.7 58.6

Source: U.S. Census Bureau

- 14. Visitors Who Travel to Foreign Countries** Construct a boxplot for the number (in millions) of visitors who traveled to a foreign country each year for a random selection of years. Comment on the skewness of the distribution.

4.3 0.5 0.6 0.8 0.5  
0.4 3.8 1.3 0.4 0.3

- 15. Areas of Islands** The data show the sizes in square miles of notable islands in the Baltic Sea and the

Aleutian Islands. Construct a boxplot for each data set and compare the distributions.

Baltic Sea	Aleutian Islands
610	275
228	1051
1154	571
1159	686
2772	350

- 16. Size of Dams** These data represent the volumes in cubic yards of the largest dams in the United States and in South America. Construct a boxplot of the data for each region and compare the distributions.

United States	South America
125,628	311,539
92,000	274,026
78,008	105,944
77,700	102,014
66,500	56,242
62,850	46,563
52,435	
50,000	

Source: New York Times Almanac.

- 17. Largest Dams** The data show the heights (in feet) of the 10 largest dams in the United States. Identify the five-number summary and the interquartile range, and draw a boxplot.

770 730 717 710 645 606 602 585 578 564

Source: U.S. Army Corps of Engineers.

- 18. Number of Tornadoes** A four-month record for the number of tornadoes in 2013–2015 is given here.

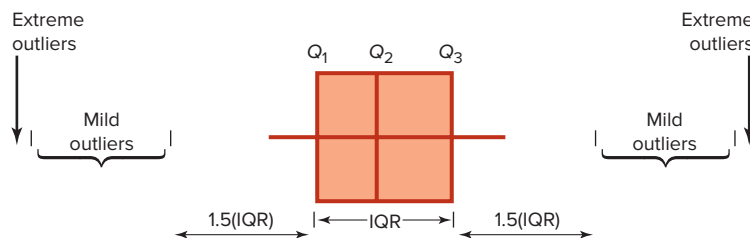
	2013	2014	2015
April	80	130	170
May	227	133	382
June	126	280	184
July	69	90	116

- Which month had the highest mean number of tornadoes for this 3-year period?
- Which year has the highest mean number of tornadoes for this 4-month period?
- Construct three boxplots and compare the distributions.

## Extending the Concepts

- 19. Unhealthy Smog Days** A *modified boxplot* can be drawn by placing a box around  $Q_1$  and  $Q_3$  and then extending the whiskers to the highest and/or lowest values within 1.5 times the interquartile range

(that is,  $Q_3 - Q_1$ ). *Mild outliers* are values greater than  $Q_3 + 1.5(IQR)$  or less than  $Q_1 - 1.5(IQR)$ . *Extreme outliers* are values greater than  $Q_3 + 3(IQR)$  or less than  $Q_1 - 3(IQR)$ .



For the data shown here, draw a modified boxplot and identify any mild or extreme outliers. The data represent the number of unhealthy smog days for a specific year for the highest 10 locations.

97 39 43 66 91  
43 54 42 53 39

Source: U.S. Public Interest Research Group and Clean Air Network.

## Technology

### TI-84 Plus

#### Step by Step

## Step by Step

### Constructing a Boxplot

To draw a boxplot:

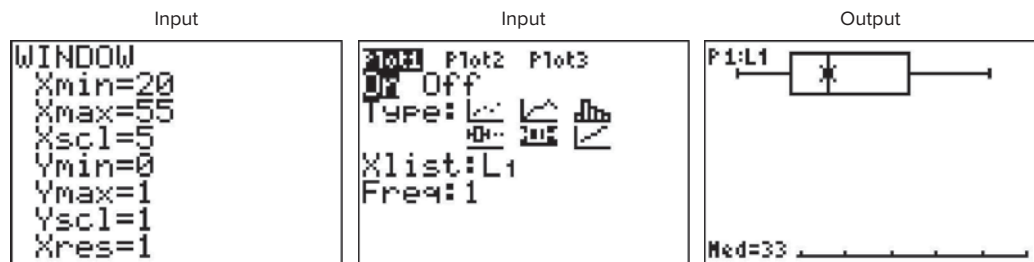
- Enter data into  $L_1$ .
- Change values in WINDOW menu, if necessary. (*Note:* Make  $X_{\min}$  somewhat smaller than the smallest data value and  $X_{\max}$  somewhat larger than the largest data value.) Change  $Y_{\min}$  to 0 and  $Y_{\max}$  to 1.
- Press [2nd] [STAT PLOT], then 1 for Plot 1.
- Press ENTER to turn on Plot 1.
- Move cursor to Boxplot symbol (fifth graph) on the Type: line, then press ENTER.
- Make sure Xlist is  $L_1$ .
- Make sure Freq is 1.
- Press GRAPH to display the boxplot.
- Press TRACE followed by  $\leftarrow$  or  $\rightarrow$  to obtain the values from the five-number summary on the boxplot.

To display two boxplots on the same screen, follow the above steps and use the 2: Plot 2 and  $L_2$  symbols.

**Example TI3-3**

Construct a boxplot for the data values:

33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31



Using the **TRACE** key along with the  $\leftarrow$  and  $\rightarrow$  keys, we obtain the five-number summary. The minimum value is 23;  $Q_1$  is 29; the median is 33;  $Q_3$  is 42; the maximum value is 51.

**EXCEL**

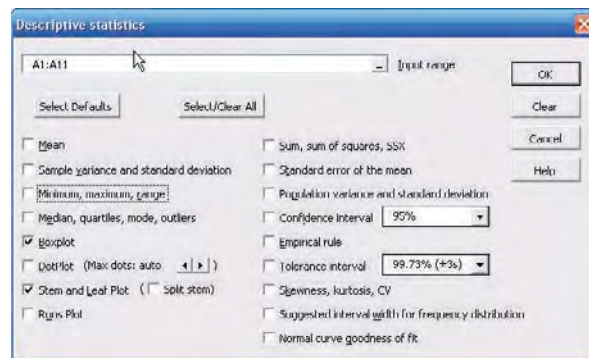
## Step by Step

**Constructing a Stem and Leaf Plot and a Boxplot****Example XL3-6**

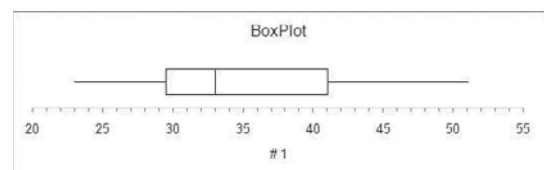
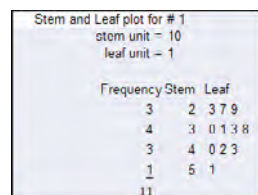
Excel does not have procedures to produce stem and leaf plots or boxplots. However, you may construct these plots by using the MegaStat Add-in available from the Online Learning Center. If you have not installed this add-in, refer to the instructions in the Excel Step by Step section of Chapter 1.

To obtain a boxplot and stem and leaf plot:

1. Enter the data values 33, 38, 43, 30, 29, 40, 51, 27, 42, 23, 31 into column A of a new Excel worksheet.
2. Select the Add-Ins tab, then MegaStat from the toolbar.
3. Select Descriptive Statistics from the MegaStat menu.
4. Enter the cell range A1:A11 in the Input range.
5. Check both Boxplot and Stem and Leaf Plot. *Note:* You may leave the other output options unchecked for this example. Click [OK].



The stem and leaf plot and the boxplot are shown below.





## MINITAB

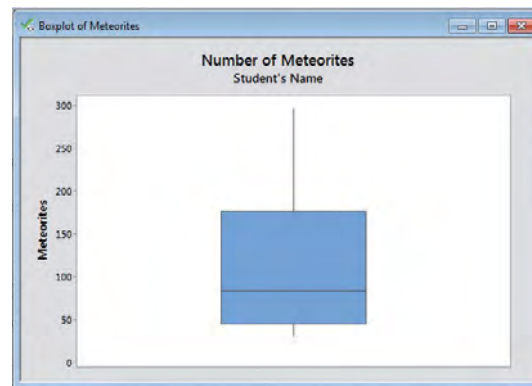
### Step by Step

#### Construct a Boxplot

1. Type in the data 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Label the column **Meteorites**.
2. Select **Graph>Boxplot**.

*Note:* Choose simple Y if all data is in one column.

3. Double-click **Meteorites** to select it for the graph variable.
4. Click on [Labels].
  - a) In the Title of the Title/Footnotes folder, type **Number of Meteorites**.
  - b) Press the [Tab] key and type **Your Name** in the text box for Subtitle 1:.
5. Click [OK] twice. The graph will be displayed in a graph window.



#### Compare Two Distributions Using Boxplots

A sample of real cheese and another sample of a cheese substitute are tested for their sodium levels in milligrams. The data are used are shown.

1. Type the data into C1 and C2 of a Minitab worksheet as shown.
  - a. Select **Graph>Boxplot** then **Multiple Y's and Simple**. **Graph>Boxplot** is an alternative menu command.

*Note:* Choose Multiple Y's when each column contains a different variable or group. Choose Simple if all data are in one column. Choose Simple with Groups if one column contains all the data and a second column identifies the group.

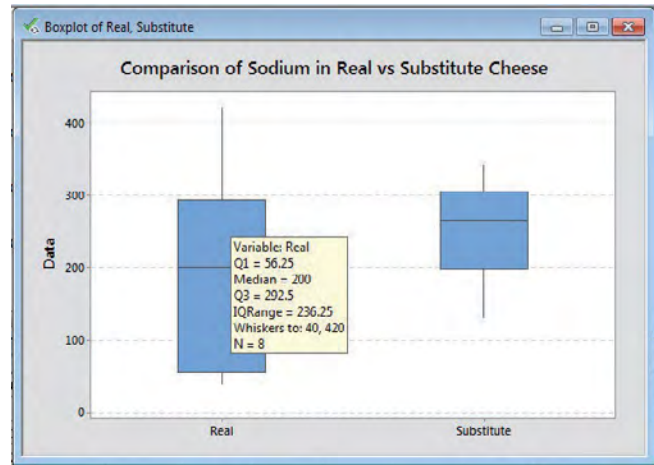
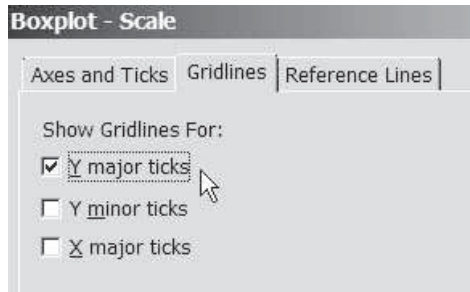
2. Drag the mouse over C1 Real and C2 Substitute, and then click the button for [Select].
3. Click on the button for [Labels]. Then in the Title dialog box type **Comparison of Sodium in Real vs Substitute Cheese** and next click [OK].

4. Click the button for [Scale].

- a. Click the tab for **Gridlines** as shown.
- b. Check the box for **Y major ticks**.

5. Click [OK] twice. Hover the mouse over the box to see the details. The quartiles may not be the same as those calculated in the text. The method varies by technology. The values will be similar.

	C1	C2
	Real	Substitute
1	310	270
2	420	180
3	45	250
4	40	290
5	220	130
6	240	260
7	180	340
8	90	310
9		



The boxplots show there is greater variation in the sodium levels for the real cheese. The median level is lower for the real cheese than the median for the sodium in the cheese substitute. The longer whisker for the real cheese indicates the distribution is skewed right. The sodium levels for the substitutes are more symmetrical and not as spread out.

The Session window will contain the descriptive statistics for sodium for each cheese type as shown. The mean amount of sodium for real cheese is 193.1 milligrams compared to 253.8 for the substitute. Even though the mean is smaller for the real group, the standard deviation is almost double, indicating greater variation in the sodium levels for the real cheese.

#### Descriptive Statistics: Sodium

Variable	Group	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Sodium	RealCheese	8	0	193.1	47.1	133.2	40.0	56.3	200.0	292.5	420.0
	Substitute	8	0	253.8	24.3	68.6	130.0	197.5	265.0	305.0	340.0

## Summary

- This chapter explains the basic ways to summarize data. These include measures of central tendency. They are the mean, median, mode, and midrange. The weighted mean can also be used. (3–1)
  - To summarize the variation of data, statisticians use measures of variation or dispersion. The three most common measures of variation are the range, variance, and standard deviation. The coefficient of variation can be used to compare the variation of two data sets. The data values are distributed according to Chebyshev's theorem or the empirical rule. (3–2)
  - There are several measures of the position of data values in the set. There are standard scores or z scores, percentiles, quartiles, and deciles. Sometimes a data set contains an extremely high or extremely low data value, called an outlier. (3–3)
  - Other methods can be used to describe a data set. These methods are the five-number summary and boxplots. These methods are called exploratory data analysis. (3–4)
- The techniques explained in Chapter 2 and this chapter are the basic techniques used in descriptive statistics.

## Important Terms

bimodal 116  
boxplot 169  
Chebyshev's theorem 139  
coefficient of variation 138  
data array 114  
decile 157  
empirical rule 141

exploratory data analysis (EDA) 168  
five-number summary 169  
interquartile range (IQR) 156  
mean 112  
median 115  
midrange 118

modal class 117  
mode 116  
multimodal 116  
negatively skewed or left-skewed distribution 122  
nonresistant statistic 157  
outlier 157

parameter 111  
percentile 149  
population mean 112  
population variance 130  
population standard deviation 130

positively skewed or right-skewed distribution 121  
 quartile 155  
 range 129

range rule of thumb 138  
 resistant statistic 157  
 sample mean 112  
 standard deviation 134

statistic 111  
 symmetric distribution 121  
 unimodal 116

variance 134  
 weighted mean 119  
 z score or standard score 148

## Important Formulas

Formula for the mean for individual data:

$$\bar{X} = \frac{\sum X}{n} \quad \mu = \frac{\sum X}{N}$$

Formula for the mean for grouped data:

$$\bar{X} = \frac{\sum f \cdot X_m}{n}$$

Formula for the weighted mean:

$$\bar{X} = \frac{\sum wX}{\sum w}$$

Formula for the midrange:

$$MR = \frac{\text{lowest value} + \text{highest value}}{2}$$

Formula for the range:

$$R = \text{highest value} - \text{lowest value}$$

Formula for the variance for population data:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Formula for the variance for sample data (shortcut formula for the unbiased estimator):

$$s^2 = \frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}$$

Formula for the variance for grouped data:

$$s^2 = \frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}$$

Formula for the standard deviation for population data:

$$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$$

Formula for the standard deviation for sample data (shortcut formula):

$$s = \sqrt{\frac{n(\sum X^2) - (\sum X)^2}{n(n-1)}}$$

Formula for the standard deviation for grouped data:

$$s = \sqrt{\frac{n(\sum f \cdot X_m^2) - (\sum f \cdot X_m)^2}{n(n-1)}}$$

Formula for the coefficient of variation:

$$CVar = \frac{s}{\bar{X}} \cdot 100 \quad \text{or} \quad CVar = \frac{\sigma}{\mu} \cdot 100$$

Range rule of thumb:

$$s \approx \frac{\text{range}}{4}$$

Expression for Chebyshev's theorem: The proportion of values from a data set that will fall within  $k$  standard deviations of the mean will be at least

$$1 - \frac{1}{k^2}$$

where  $k$  is a number greater than 1.

Formula for the z score (standard score):

$$z = \frac{X - \bar{X}}{s} \quad \text{or} \quad z = \frac{X - \mu}{\sigma}$$

Formula for the cumulative percentage:

$$\text{Cumulative \%} = \frac{\text{cumulative frequency}}{n} \cdot 100$$

Formula for the percentile rank of a value  $X$ :

$$\text{Percentile} = \frac{\text{number of values below } X + 0.5}{\text{total number of values}} \cdot 100$$

Formula for finding a value corresponding to a given percentile:

$$c = \frac{n \cdot p}{100}$$

Formula for interquartile range:

$$IQR = Q_3 - Q_1$$

## Review Exercises

### SECTION 3-1

1. **Bank Failures** The data show the number of bank failures over a 15-year period. Find the mean, median, midrange, and mode for the data.

14 24 51 92 157 148 30 3  
 0 0 4 3 11 4 7

Source: Federal Deposit Insurance Corporation.

2. **Shark Attacks** The number of shark attacks and deaths over a recent 5-year period is shown. Find the mean, median, mode, and midrange for the data.

Attacks	71	64	61	65	57
Deaths	1	4	4	7	4

- 3. Systolic Blood Pressure** The data show the systolic blood pressure of 30 college students. Find the mean and modal class.

Class	Frequency
105–109	2
110–114	5
115–119	6
120–124	8
125–129	8
130–134	1

- 4. SAT Scores** The mean SAT math scores for selected states are represented. Find the mean class and modal class.

Score	Frequency
478–504	4
505–531	6
532–558	2
559–585	2
586–612	2

Source: World Almanac.

- 5. Households of Four Television Networks** A survey showed the number of viewers and number of households of four television networks. Find the average number of viewers, using the weighted mean.

<b>Households</b>	1.4	0.8	0.3	1.6
<b>Viewers (in millions)</b>	1.6	0.8	0.4	1.8

Source: Nielsen Media Research.

- 6. Investment Earnings** An investor calculated these percentages of each of three stock investments with payoffs as shown. Find the average payoff. Use the weighted mean.

Stock	Percent	Payoff
A	30	\$10,000
B	50	3,000
C	20	1,000

## SECTION 3–2

- 7. Confirmed Measles Cases** The data show a sample of the number of confirmed measles cases over a recent 12-year period. Find the range, variance, and standard deviation for the data.

212	63	71	140	43	55	66	37	56
44	116	86						

Source: Centers for Disease Control and Prevention.

- 8. Tallest Buildings** The number of stories in the 13 tallest buildings in Houston are shown. Find the range, variance, and standard deviation for the data.

75	71	64	56	53	55	47	55	52	50
50	50	47							

Source: World Almanac.

- 9. Rise in Tides** Shown here is a frequency distribution for the rise in tides at 30 selected locations in the United States. Find the variance and standard deviation for the data.

Rise in tides (inches)	Frequency
12.5–27.5	6
27.5–42.5	3
42.5–57.5	5
57.5–72.5	8
72.5–87.5	6
87.5–102.5	2

- 10. Fuel Capacity** The fuel capacity in gallons for randomly selected cars is shown here. Find the variance and standard deviation for the data.

Class	Frequency
10–12	6
13–15	4
16–18	14
19–21	15
22–24	8
25–27	2
28–30	1
	<u>50</u>

- 11.** If the range of a data set is 24, find the approximate value of the standard deviation, using the range rule of thumb.
- 12.** If the range of a data set is 56, find the approximate value of the standard deviation, using the range rule of thumb.
- 13. Textbooks in Professors' Offices** If the average number of textbooks in professors' offices is 16, the standard deviation is 5, and the average age of the professors is 43, with a standard deviation of 8, which data set is more variable?
- 14. Magazines in Bookstores** A survey of bookstores showed that the average number of magazines carried is 56, with a standard deviation of 12. The same survey showed that the average length of time each store had been in business was 6 years, with a standard deviation of 2.5 years. Which is more variable, the number of magazines or the number of years?
- 15. Cost of Car Rentals** A survey of car rental agencies shows that the average cost of a car rental is \$0.32 per mile. The standard deviation is \$0.03. Using Chebyshev's theorem, find the range in which at least 75% of the data values will fall.

- 16. Average Earnings of Workers** The average earnings of year-round full-time workers 25–34 years old with a bachelor's degree or higher were \$58,500 in 2003. If the standard deviation is \$11,200, what can you say about the percentage of these workers who earn.

- Between \$47,300 and \$69,700?
- More than \$80,900?
- How likely is it that someone earns more than \$100,000?

Source: *New York Times Almanac*.

- 17. Labor Charges** The average labor charge for automobile mechanics is \$54 per hour. The standard deviation is \$4. Find the minimum percentage of data values that will fall within the range of \$48 to \$60. Use Chebyshev's theorem.

- 18. Costs to Train Employees** For a certain type of job, it costs a company an average of \$231 to train an employee to perform the task. The standard deviation is \$5. Find the minimum percentage of data values that will fall in the range of \$219 to \$243. Use Chebyshev's theorem.

- 19. Cost of a Man's Haircut** The average cost of a man's haircut is \$21. The standard deviation is \$4. If the variable is approximately bell-shaped, within what limits would 68% of the haircut cost?

- 20. Exam Completion Time** The mean time it takes a group of students to complete a statistics final exam is 44 minutes, and the standard deviation is 9 minutes. Within what limits would you expect approximately 95% of the students to complete the exam? Assume the variable is approximately normally distributed.

- 21. Cases of Meningitis** The data show the number of specific recorded cases of meningitis for 14 specific states.
- |    |    |   |    |    |    |   |   |   |   |   |
|----|----|---|----|----|----|---|---|---|---|---|
| 10 | 1  | 1 | 28 | 15 | 41 | 4 | 4 | 8 | 2 | 3 |
| 53 | 34 | 1 |    |    |    |   |   |   |   |   |

Source: *Centers for Disease Control and Prevention*

Find the  $z$  values for each

- 10
- 28
- 41

### SECTION 3–3

- 22. Exam Grades** Which of these exam grades has a better relative position?

- A grade of 82 on a test with  $\bar{X} = 85$  and  $s = 6$
- A grade of 56 on a test with  $\bar{X} = 60$  and  $s = 5$

- 23.** The number of police calls a small police department received each month is shown in the frequency distribution.

Class limits	Frequency
39.9–42.8	2
42.9–45.8	2
45.9–48.8	5
48.9–51.8	5
51.9–54.8	12
54.9–57.8	3

- Construct a percentile graph.
- Find the values that correspond to the 35th, 65th, and 85th percentiles.
- Find the percentile of values 44, 48, and 54.

- 24. Printer Repairs** The frequency distribution shows the number of days it took to fix each of 80 computer's printers.

Class limits	Frequency
1–3	7
4–6	9
7–9	32
10–12	20
13–15	12
	<hr/> 80

- Construct a percentile graph.
- Find the 20th, 50th, and 70th percentiles.
- Find the percentile values of 5, 10, and 14.

- 25.** Check each data set for outliers.

- 506, 511, 517, 514, 400, 521
- 3, 7, 9, 6, 8, 10, 14, 16, 20, 12

- 26.** Check each data set for outliers.

- 14, 18, 27, 26, 19, 13, 5, 25
- 112, 157, 192, 116, 153, 129, 131

### SECTION 3–4

- 27. Named Storms** The data show the number of named storms for the years 1851–1860 and 1941 and 1950. Construct a boxplot for each data set and compare the distributions.

<b>1851–1860</b>	6	5	8	5	5	6	4	6	8	7
<b>1941–1950</b>	6	10	10	11	11	6	9	9	13	13

Source: *National Hurricane Center*

- 28. Hours Worked** The data shown here represent the number of hours that 12 part-time employees at a toy store worked during the weeks before and after Christmas. Construct two boxplots and compare the distributions.

<b>Before</b>	38	16	18	24	12	30	35	32	31	30	24	35
<b>After</b>	26	15	12	18	24	32	14	18	16	18	22	12

## STATISTICS TODAY

### How Long Are You Delayed by Road Congestion? — Revisited

The average number of hours per year that a driver is delayed by road congestion is listed here.

Los Angeles	56
Atlanta	53
Seattle	53
Houston	50
Dallas	46
Washington	46
Austin	45
Denver	45
St. Louis	44
Orlando	42
U.S. average	36

Source: Texas Transportation Institute.

By making comparisons using averages, you can see that drivers in these 10 cities are delayed by road congestion more than the national average.

## Data Analysis

A Data Bank is found in Appendix B, or on the World Wide Web by following links from [www.mhhe.com/math/stat/bluman/](http://www.mhhe.com/math/stat/bluman/)

- From the Data Bank, choose one of the following variables: age, weight, cholesterol level, systolic pressure, IQ, or sodium level. Select at least 30 values, and find the mean, median, mode, and midrange. State which measurement of central tendency best describes the average and why.
- Find the range, variance, and standard deviation for the data selected in Exercise 1.
- From the Data Bank, choose 10 values from any variable, construct a boxplot, and interpret the results.
- Randomly select 10 values from the number of suspensions in the local school districts in southwestern Pennsylvania in Data Set V in Appendix B. Find the mean, median, mode, range, variance, and standard deviation of the number of suspensions by using the Pearson coefficient of skewness.
- Using the data from Data Set VII in Appendix B, find the mean, median, mode, range, variance, and standard deviation of the acreage owned by the municipalities. Comment on the skewness of the data, using the Pearson coefficient of skewness.

## Chapter Quiz

Determine whether each statement is true or false. If the statement is false, explain why.

- When the mean is computed for individual data, all values in the data set are used.
- The mean cannot be found for grouped data when there is an open class.
- A single, extremely large value can affect the median more than the mean.
- One-half of all the data values will fall above the mode, and one-half will fall below the mode.
- In a data set, the mode will always be unique.
- The range and midrange are both measures of variation.
- One disadvantage of the median is that it is not unique.
- The mode and midrange are both measures of variation.
- If a person's score on an exam corresponds to the 75th percentile, then that person obtained 75 correct answers out of 100 questions.



**Select the best answer.**

- 10.** What is the value of the mode when all values in the data set are different?
- a. 0                      c. There is no mode.  
b. 1                      d. It cannot be determined unless the data values are given.
- 11.** When data are categorized as, for example, places of residence (rural, suburban, urban), the most appropriate measure of central tendency is the
- a. Mean                  c. Mode  
b. Median                d. Midrange
- 12.**  $P_{50}$  corresponds to
- a.  $Q_2$                     c. IQR  
b.  $D_5$                     d. Midrange
- 13.** Which is not part of the five-number summary?
- a.  $Q_1$  and  $Q_3$           c. The median  
b. The mean              d. The smallest and the largest data values
- 14.** A statistic that tells the number of standard deviations a data value is above or below the mean is called
- a. A quartile            c. A coefficient of variation  
b. A percentile          d. A z score
- 15.** When a distribution is bell-shaped, approximately what percentage of data values will fall within 1 standard deviation of the mean?
- a. 50%                    c. 95%  
b. 68%                    d. 99.7%

**Complete these statements with the best answer.**

- 16.** A measure obtained from sample data is called a(n) \_\_\_\_\_.
- 17.** Generally, Greek letters are used to represent \_\_\_\_\_, and Roman letters are used to represent \_\_\_\_\_.
- 18.** The positive square root of the variance is called the \_\_\_\_\_.
- 19.** The symbol for the population standard deviation is \_\_\_\_\_.
- 20.** When the sum of the lowest data value and the highest data value is divided by 2, the measure is called the \_\_\_\_\_.
- 21.** If the mode is to the left of the median and the mean is to the right of the median, then the distribution is \_\_\_\_\_ skewed.
- 22.** An extremely high or extremely low data value is called a(n) \_\_\_\_\_.
- 23. Miles per Gallon** The number of highway miles per gallon of the 10 worst vehicles is shown.
- 12 15 13 14 15 16 17 16 17 18

Source: Pittsburgh Post Gazette.

Find each of these.

- a. Mean                      e. Range  
b. Median                    f. Variance  
c. Mode                      g. Standard deviation  
d. Midrange

- 24. Errors on a Typing Test** The distribution of the number of errors that 10 students made on a typing test is shown.

Errors	Frequency
0–2	1
3–5	3
6–8	4
9–11	1
12–14	1

Find each of these.

- a. Mean                      c. Variance  
b. Modal class              d. Standard deviation

- 25. Employee Years of Service** In an advertisement, a retail store stated that its employees averaged 9 years of service. The distribution is shown here.

Number of employees	Years of service
8	2
2	6
3	10

Using the weighted mean, calculate the correct average.

- 26. Newspapers for Sale** The average number of newspapers for sale in an airport newsstand is 56 with a standard deviation of 6. The average number of newspapers for sale in a convenience store is 44 with a standard deviation of 5. Which data set is more variable?

- 27. Delivery Charges** The average delivery charge for a refrigerator is \$32. The standard deviation is \$4. Find the minimum percentage of data values that will fall in the range of \$20 to \$44. Use Chebyshev's theorem.

- 28. SAT Scores** The average national SAT score is 1019. If we assume a bell-shaped distribution and a standard deviation equal to 110, what percentage of scores will you expect to fall above 1129? Above 799?

Source: New York Times Almanac.

- 29.** If the range of a data set is 18, estimate the standard deviation of the data.
- 30. Test Scores** A student scored 76 on a general science test where the class mean and standard deviation were 82 and 8, respectively; he also scored 53 on a psychology test where the class mean and standard deviation were 58 and 3, respectively. In which class was his relative position higher?
- 31. Exam Scores** On a philosophy comprehensive exam, this distribution was obtained from 25 students.



Score	Frequency
40.5–45.5	3
45.5–50.5	8
50.5–55.5	10
55.5–60.5	3
60.5–65.5	1

- Construct a percentile graph.
- Find the values that correspond to the 22nd, 78th, and 99th percentiles.
- Find the percentiles of the values 52, 43, and 64.

- 32. Gas Prices for Rental Cars** The first column of these data represents the prebuy gas price of a rental car, and the second column represents the price charged if the car is returned without refilling the gas tank for a

selected car rental company. Draw two boxplots for the data and compare the distributions. (Note: The data were collected several years ago.)

Prebuy cost	No prebuy cost
\$1.55	\$3.80
1.54	3.99
1.62	3.99
1.65	3.85
1.72	3.99
1.63	3.95
1.65	3.94
1.72	4.19
1.45	3.84
1.52	3.94

Source: USA TODAY.

## Critical Thinking Challenges

- 1. Average Cost of Weddings** Averages give us information to help us see where we stand and enable us to make comparisons. Here is a study on the average cost of a wedding. What type of average—mean, median, mode, or midrange—might have been used for each category?

### Wedding Costs

The average cost of a wedding varies each year. The average cost of a wedding in 2015 was \$26,444. The average cost of the reception venue, catering, and rentals were \$11,784. The cost for wedding flowers was \$1579. The cost of the wedding invitations was \$896.

The cost of a wedding planner was \$1646. The cost of photography and videography was \$1646.

- 2. Average Cost of Smoking** The average yearly cost of smoking a pack of cigarettes a day is \$1190. Find the average cost of a pack of cigarettes in your area, and compute the cost per day for 1 year. Compare your answer with the one in the article.
- 3. Ages of U.S. Residents** The table shows the median ages of residents for the 10 oldest states and the 10 youngest states of the United States including Washington, D.C. Explain why the median is used instead of the mean.

10 Oldest			10 Youngest		
Rank	State	Median age	Rank	State	Median age
1	West Virginia	38.9	51	Utah	27.1
2	Florida	38.7	50	Texas	32.3
3	Maine	38.6	49	Alaska	32.4
4	Pennsylvania	38.0	48	Idaho	33.2
5	Vermont	37.7	47	California	33.3
6	Montana	37.5	46	Georgia	33.4
7	Connecticut	37.4	45	Mississippi	33.8
8	New Hampshire	37.1	44	Louisiana	34.0
9	New Jersey	36.7	43	Arizona	34.2
10	Rhode Island	36.7	42	Colorado	34.3

Source: U.S. Census Bureau.

## Data Projects

Where appropriate, use MINITAB, the TI-84 Plus, Excel, or a computer program of your choice to complete the following exercises.

- 1. Business and Finance** Use the data collected in data project 1 of Chapter 2 regarding earnings per share. Determine the mean, mode, median, and midrange for the two data sets. Is one measure of center more appropriate than the other for these data? Do the measures of center appear similar? What does this say about the symmetry of the distribution?
- 2. Sports and Leisure** Use the data collected in data project 2 of Chapter 2 regarding home runs. Determine the mean, mode, median, and midrange for the two data sets. Is one measure of center more appropriate than the other for these data? Do the measures of center appear similar? What does this say about the symmetry of the distribution?
- 3. Technology** Use the data collected in data project 3 of Chapter 2. Determine the mean for the frequency table created in that project. Find the actual mean length of all 50 songs. How does the grouped mean compare to the actual mean?
- 4. Health and Wellness** Use the data collected in data project 6 of Chapter 2 regarding heart rates. Determine

the mean and standard deviation for each set of data. Do the means seem very different from one another? Do the standard deviations appear very different from one another?

- Politics and Economics** Use the data collected in data project 5 of Chapter 2 regarding delegates. Use the formulas for population mean and standard deviation to compute the parameters for all 50 states. What is the  $z$  score associated with California? Delaware? Ohio?

Which states are more than 2 standard deviations from the mean?

- Your Class** Use your class as a sample. Determine the mean, median, and standard deviation for the age of students in your class. What  $z$  score would a 40-year-old have? Would it be unusual to have an age of 40? Determine the skew of the data, using the Pearson coefficient of skewness. (See Exercise 48 in Exercise 3–2.)

## Answers to Applying the Concepts

### Section 3–1 Teacher Salaries

- The sample mean is \$22,921.67, the sample median is \$16,500, and the sample mode is \$11,000. If you work for the school board and do not want to raise salaries, you could say that the average teacher salary is \$22,921.67.
- If you work for the teachers' union and want a raise for the teachers, the sample mode of \$11,000 would be a good measure of center to report.
- The outlier is \$107,000. With the outlier removed, the sample mean is \$15,278.18, the sample median is \$16,400, and the sample mode is still \$11,000. The mean is greatly affected by the outlier and allows the school board to report an average teacher salary that is not representative of a "typical" teacher salary.
- If the salaries represented every teacher in the school district, the averages would be parameters, since we have data from the entire population.
- The mean and midrange can be misleading in the presence of outliers, since they are greatly affected by these extreme values.
- Since the mean is greater than both the median and the mode, the distribution is skewed to the right (positively skewed).

### Section 3–2 Blood Pressure

- Chebyshev's theorem does not work for scores within 1 standard deviation of the mean.
- At least 75% (900) of the normotensive men have systolic blood pressures that fall between 105 and 141 mm Hg.
- About 68% (952) of the normotensive women have diastolic blood pressures between 69 and 83 mm Hg. About 95% (1330) of the normotensive women have diastolic blood pressures between 62 and 90 mm Hg. About 99.7% (1396) of the normotensive women have diastolic blood pressures between 55 and 97 mm Hg. About 68% (884) of the hypertensive women have diastolic blood pressures between 78 and 98 mm Hg. About 95% (1235) of the hypertensive women have diastolic blood pressures between 68 and 108 mm Hg. About 99.7% (1296) of the hypertensive women have diastolic blood pressures between 58 and 118 mm Hg.
- About 68% (816) of the normotensive men have systolic blood pressures between 114 and 132 mm Hg. About 95% (1140) of the normotensive men have systolic blood pressures between 105 and 141 mm Hg. About

99.7% (1196) of the normotensive men have systolic blood pressures between 96 and 150 mm Hg.

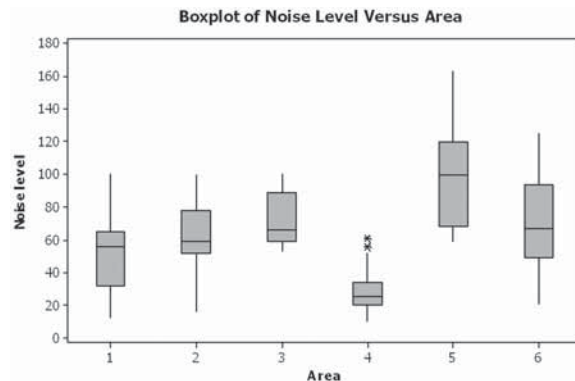
About 68% (748) of the hypertensive men have systolic blood pressures between 136 and 170 mm Hg. About 95% (1045) of the hypertensive men have systolic blood pressures between 119 and 187 mm Hg. About 99.7% (1097) of the hypertensive men have systolic blood pressures between 102 and 204 mm Hg.

The "68%" ranges do not overlap. The "95%" and "99.7%" ranges overlap.

### Section 3–3 Determining Dosages

- Standard scores ( $z$ -scores), percentiles, deciles, and quartiles can be used to describe the data results.
- Since there are 10 mice in the upper quartile, this would mean that 4 of them survived.
- The percentiles would give us the position of a single mouse with respect to all other mice.
- The quartiles divide the data into four groups of equal size.
- Standard scores would give us the position of a single mouse with respect to the mean time until the onset of sepsis.

### Section 3–4 The Noisy Workplace



From this boxplot, we see that about 25% of the readings in area 5 are above the safe hearing level of 120 decibels. Those workers in area 5 should definitely have protective ear wear. One of the readings in area 6 is above the safe hearing level. It might be a good idea to provide protective ear wear to those workers in area 6 as well. Areas 1–4 appear to be "safe" with respect to hearing level, with area 4 being the safest.

# Chapter 15

## Exploratory Data Analysis

Matthieu Komorowski, Dominic C. Marshall, Justin D. Saliccioli  
and Yves Crutain

### Learning Objectives

- Why is EDA important during the initial exploration of a dataset?
- What are the most essential tools of graphical and non-graphical EDA?

## 15.1 Introduction

Exploratory data analysis (EDA) is an essential step in any research analysis. The primary aim with exploratory analysis is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis. It also provides tools for hypothesis generation by visualizing and understanding the data usually through graphical representation [1]. EDA aims to assist the natural patterns recognition of the analyst. Finally, feature selection techniques often fall into EDA. Since the seminal work of Tukey in 1977, EDA has gained a large following as the gold standard methodology to analyze a data set [2, 3]. According to Howard Seltman (Carnegie Mellon University), “loosely speaking, any method of looking at data that does not include formal statistical modeling and inference falls under the term exploratory data analysis” [4].

EDA is a fundamental early step after data collection (see Chap. 11) and pre-processing (see Chap. 12), where the data is simply visualized, plotted, manipulated, without any assumptions, in order to help assessing the quality of the data and building models. “Most EDA techniques are graphical in nature with a few quantitative techniques. The reason for the heavy reliance on graphics is that by its very nature the main role of EDA is to explore, and graphics gives the analysts unparalleled power to do so, while being ready to gain insight into the data. There are many ways to categorize the many EDA techniques” [5].

---

**Electronic supplementary material** The online version of this chapter (doi:[10.1007/978-3-319-43742-2\\_15](https://doi.org/10.1007/978-3-319-43742-2_15)) contains supplementary material, which is available to authorized users.

The interested reader will find further information in the textbooks of Hill and Lewicki [6] or the NIST/SEMATECH e-Handbook [1]. Relevant R packages are available on the CRAN website [7].

The objectives of EDA can be summarized as follows:

1. Maximize insight into the database/understand the database structure;
2. Visualize potential relationships (direction and magnitude) between exposure and outcome variables;
3. Detect outliers and anomalies (values that are significantly different from the other observations);
4. Develop parsimonious models (a predictive or explanatory model that performs with as few exposure variables as possible) or preliminary selection of appropriate models;
5. Extract and create clinically relevant variables.

EDA methods can be cross-classified as:

- Graphical or non-graphical methods
- Univariate (only one variable, exposure or outcome) or multivariate (several exposure variables alone or with an outcome variable) methods.

15.2 Part 1—Theoretical Concepts

15.2.1 Suggested EDA Techniques

Tables 15.1 and 15.2 suggest a few EDA techniques depending on the type of data and the objective of the analysis.

**Table 15.1** Suggested EDA techniques depending on the type of data

Type of data	Suggested EDA techniques
Categorical	Descriptive statistics
Univariate continuous	Line plot, Histograms
Bivariate continuous	2D scatter plots
2D arrays	Heatmap
Multivariate: trivariate	3D scatter plot or 2D scatter plot with a 3rd variable represented in different color, shape or size
Multiple groups	Side-by-side boxplot

**Table 15.2** Most useful EDA techniques depending on the objective

Objective	Suggested EDA techniques
Getting an idea of the distribution of a variable	Histogram
Finding outliers	Histogram, scatterplots, box-and-whisker plots
Quantify the relationship between two variables (one exposure and one outcome)	2D scatter plot +/-curve fitting Covariance and correlation
Visualize the relationship between two exposure variables and one outcome variable	Heatmap
Visualization of high-dimensional data	t-SNE or PCA + 2D/3D scatterplot

*t-SNE* t-distributed stochastic neighbor embedding, *PCA* Principal component analysis

**Table 15.3** Example of tabulation table

	Group count	Frequency (%)
Green ball	15	75
Red ball	5	25
Total	20	100

15.2.2 Non-graphical EDA

These non-graphical methods will provide insight into the characteristics and the distribution of the variable(s) of interest.

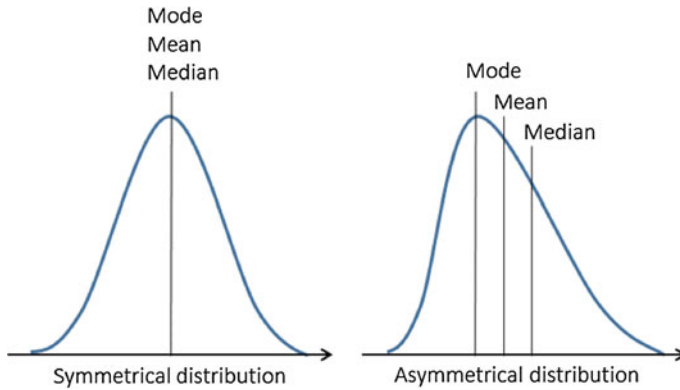
Univariate Non-graphical EDA

*Tabulation of Categorical Data (Tabulation of the Frequency of Each Category)*

A simple univariate non-graphical EDA method for categorical variables is to build a table containing the count and the fraction (or frequency) of data of each category. An example of tabulation is shown in the case study (Table 15.3).

*Characteristics of Quantitative Data: Central Tendency, Spread, Shape of the Distribution (Skewness, Kurtosis)*

Sample statistics express the characteristics of a sample using a limited set of parameters. They are generally seen as estimates of the corresponding population parameters from which the sample comes from. These characteristics can express the central tendency of the data (arithmetic mean, median, mode), its spread (variance, standard deviation, interquartile range, maximum and minimum value) or some features of its distribution (skewness, kurtosis). Many of those characteristics can easily be seen qualitatively on a histogram (see below). Note that these characteristics can only be used for quantitative variables (not categorical).



**Fig. 15.1** Symmetrical versus asymmetrical (skewed) distribution, showing mode, mean and median

### Central tendency parameters

The arithmetic mean, or simply called the mean is the sum of all data divided by the number of values. The median is the middle value in a list containing all the values sorted. Because the median is affected little by extreme values and outliers, it is said to be more “robust” than the mean (Fig. 15.1).

### Variance

When calculated on the entirety of the data of a population (which rarely occurs), the variance  $\sigma^2$  is obtained by dividing the sum of squares by  $n$ , the size of the population.

The sample formula for the variance of observed data conventionally has  $n-1$  in the denominator instead of  $n$  to achieve the property of “unbiasedness”, which roughly means that when calculated for many different random samples from the same population, the average should match the corresponding population quantity (here  $\sigma^2$ ).  $s^2$  is an unbiased estimator of the population variance  $\sigma^2$ .

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(n - 1)} \quad (15.1)$$

The standard deviation is simply the square root of the variance. Therefore it has the same units as the original data, which helps make it more interpretable.

The sample standard deviation is usually represented by the symbol  $s$ . For a theoretical Gaussian distribution, mean plus or minus 1, 2 or 3 standard deviations holds 68.3, 95.4 and 99.7 % of the probability density, respectively.

### Interquartile range (IQR)

The IQR is calculated using the boundaries of data situated between the 1st and the 3rd quartiles. Please refer to the Chap. 13 “Noise versus Outliers” for further detail about the IQR.

$$IQR = Q_3 - Q_1 \quad (15.2)$$

In the same way that the median is more robust than the mean, the IQR is a more robust measure of spread than variance and standard deviation and should therefore be preferred for small or asymmetrical distributions.

### Important rule:

- **Symmetrical distribution** (not necessarily normal) **and N > 30**: express results as mean  $\pm$  standard deviation.
- **Asymmetrical distribution or N < 30 or evidence for outliers**: use median  $\pm$  IQR, which are more robust.

### Skewness/kurtosis

Skewness is a measure of a distribution’s asymmetry. Kurtosis is a summary statistic communicating information about the tails (the smallest and largest values) of the distribution. Both quantities can be used as a means to communicate information about the distribution of the data when graphical methods cannot be used. More information about these quantities can be found in [9]).

### Summary

We provide as a reference some of the common functions in R language for generating summary statistics relating to measures of central tendency (Table 15.4).

### Testing the Distribution

Several non-graphical methods exist to assess the normality of a data set (whether it was sampled from a normal distribution), like the Shapiro-Wilk test for example. Please refer to the function called “Distribution” in the GitHub repository for this book (see code appendix at the end of this Chapter).

**Table 15.4** Main R functions for basic measure of central tendencies and variability

Function	Description
summary(x)	General description of a vector
max(x)	Maximum value
mean(x)	Average or mean value
median(x)	Median value
min(x)	Smallest value
sd(x)	Standard deviation
var(x)	Variance, measure the spread or dispersion of the values
IQR(x)	Interquartile range



### *Finding Outliers*

Several statistical methods for outlier detection fall into EDA techniques, like Tukey’s method, Z-score, studentized residuals, etc [8]. Please refer to the Chap. 14 “Noise versus Outliers” for more detail about this topic.

### **Multivariate Non-graphical EDA**

#### *Cross-Tabulation*

Cross-tabulation represents the basic bivariate non-graphical EDA technique. It is an extension of tabulation that works for categorical data and quantitative data with only a few variables. For two variables, build a two-way table with column headings matching the levels of one variable and row headings matching the levels of the other variable, then fill in the counts of all subjects that share a pair of levels. The two variables may be both exposure, both outcome variables, or one of each.

#### *Covariance and Correlation*

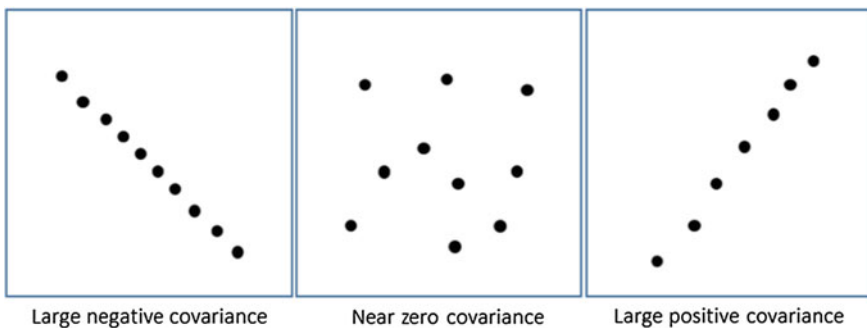
Covariance and correlation measure the degree of the relationship between two random variables and express how much they change together (Fig. 15.2).

The covariance is computed as follows:

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (15.3)$$

where  $x$  and  $y$  are the variables,  $n$  the number of data points in the sample,  $\bar{x}$  the mean of the variable  $x$  and  $\bar{y}$  the mean of the variable  $y$ .

A positive covariance means the variables are positively related (they move together in the same direction), while a negative covariance means the variables are inversely related. A problem with covariance is that its value depends on the scale of the values of the random variables. The larger the values of  $x$  and  $y$ , the larger the



**Fig. 15.2** Examples of covariance for three different data sets

covariance. It makes it impossible for example to compare covariances from data sets with different scales (e.g. pounds and inches). This issue can be fixed by dividing the covariance by the product of the standard deviation of each random variable, which gives Pearson’s correlation coefficient.

Correlation is therefore a scaled version of covariance, used to assess the linear relationship between two variables and is calculated using the formula below.

$$Cor(x, y) = \frac{Cov(x, y)}{s_x s_y} \quad (15.4)$$

where  $Cov(x, y)$  is the covariance between  $x$  and  $y$  and  $s_x, s_y$  are the sample standard deviations of  $x$  and  $y$ .

The significance of the correlation coefficient between two normally distributed variables can be evaluated using Fisher’s  $z$  transformation (see the `cor.test` function in R for more details). Other tests exist for measuring the non-parametric relationship between two variables, such as Spearman’s  $\rho$  or Kendall’s  $\tau$ .

### 15.2.3 Graphical EDA

#### Univariate Graphical EDA

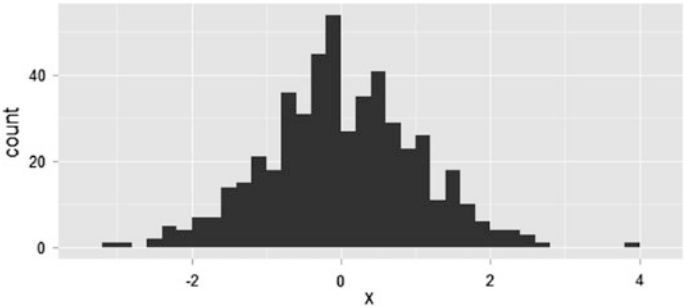
##### Histograms

Histograms are among the most useful EDA techniques, and allow you to gain insight into your data, including distribution, central tendency, spread, modality and outliers.

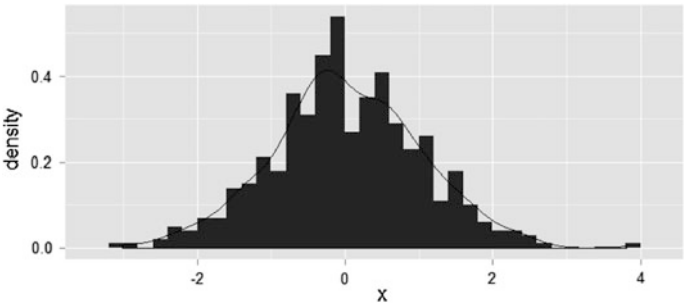
Histograms are bar plots of counts versus subgroups of an exposure variable. Each bar represents the frequency (count) or proportion (count divided by total count) of cases for a range of values. The range of data for each bar is called a bin. Histograms give an immediate impression of the shape of the distribution (symmetrical, uni/multimodal, skewed, outliers...). The number of bins heavily influences the final aspect of the histogram; a good practice is to try different values, generally from 10 to 50. Some examples of histograms are shown below as well as in the case studies. Please refer to the function called “Density” in the GitHub repository for this book (see code appendix at the end of this Chapter) (Figs. 15.3 and 15.4).

Histograms enable to confirm that an operation on data was successful. For example, if you need to log-transform a data set, it is interesting to plot the histogram of the distribution of the data before and after the operation (Fig. 15.5).

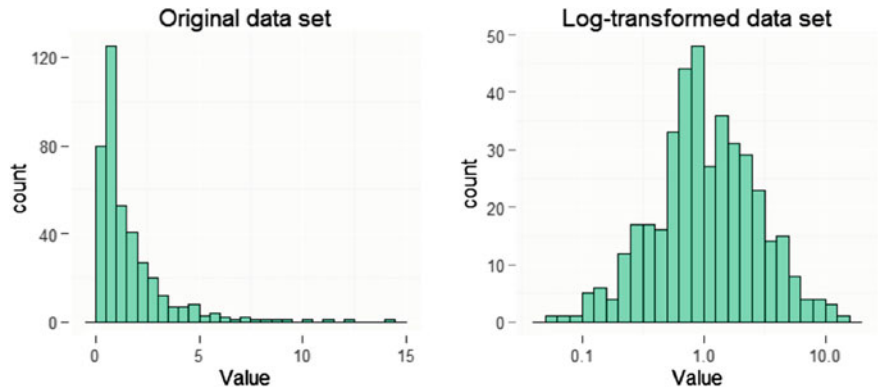
Histograms are interesting for finding outliers. For example, pulse oximetry can be expressed in fractions (range between 0 and 1) or percentage, in medical records. Figure 15.6 is an example of a histogram showing the distribution of pulse oximetry, clearly showing the presence of outliers expressed in a fraction rather than as a percentage.



**Fig. 15.3** Example of histogram



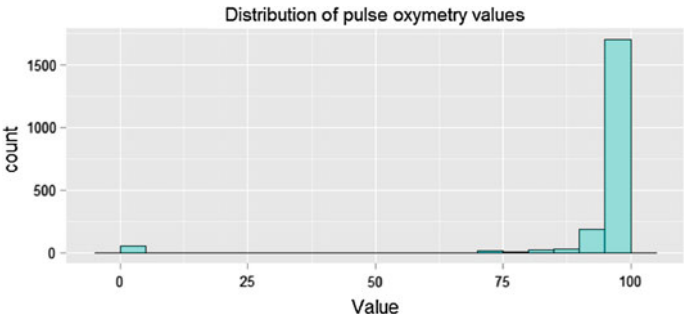
**Fig. 15.4** Example of histogram with density estimate



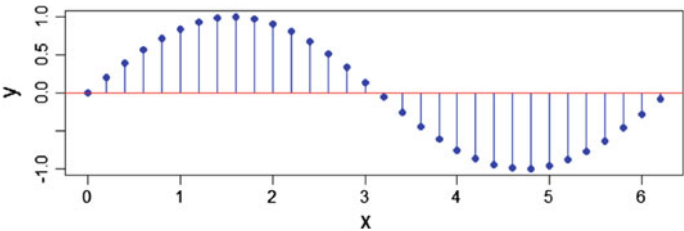
**Fig. 15.5** Example of the effect of a log transformation on the distribution of the dataset

*Stem Plots*

Stem and leaf plots (also called stem plots) are a simple substitution for histograms. They show all data values and the shape of the distribution. For an example, Please refer to the function called “Stem Plot” in the GitHub repository for this book (see code appendix at the end of this Chapter) (Fig. 15.7).



**Fig. 15.6** Distribution of pulse oximetry



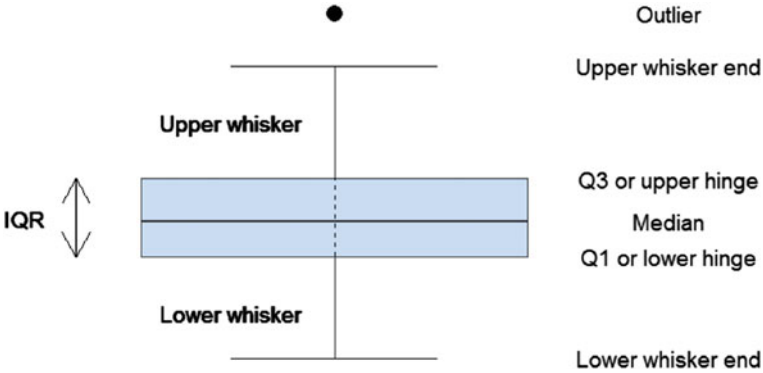
**Fig. 15.7** Example of stem plot

*Boxplots*

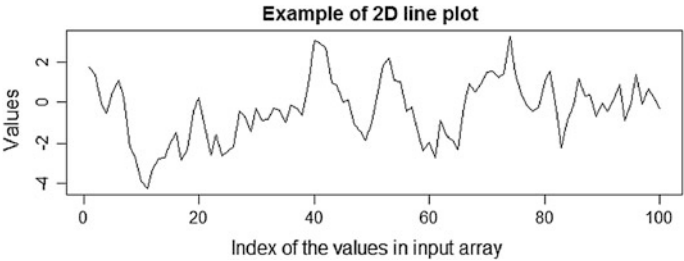
Boxplots are interesting for representing information about the central tendency, symmetry, skew and outliers, but they can hide some aspects of the data such as multimodality. Boxplots are an excellent EDA technique because they rely on robust statistics like median and IQR.

Figure 15.8 shows an annotated boxplot which explains how it is constructed. The central rectangle is limited by Q1 and Q3, with the middle line representing the median of the data. The whiskers are drawn, in each direction, to the most extreme point that is less than 1.5 IQR beyond the corresponding hinge. Values beyond 1.5 IQR are considered outliers.

The “outliers” identified by a boxplot, which could be called “boxplot outliers” are defined as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1. This does not by itself indicate a problem with those data points. Boxplots are an exploratory technique, and you should consider designation as a boxplot outlier as just a suggestion that the points might be mistakes or otherwise unusual. Also, points not designated as boxplot outliers may also be mistakes. It is also important to realize that the number of boxplot outliers depends strongly on the size of the sample. In fact, for data that is perfectly normally distributed, we expect 0.70 % (about 1 in 140 cases) to be “boxplot outliers”, with approximately half in either direction.



**Fig. 15.8** Example of boxplot with annotations



**Fig. 15.9** Example of 2D line plot

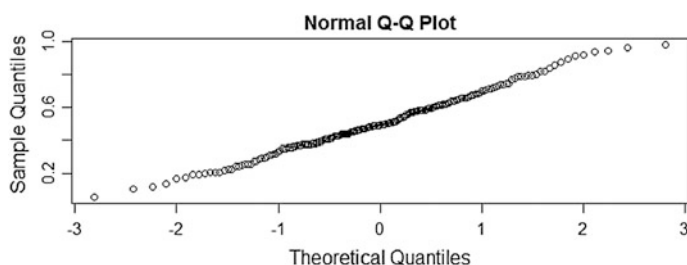
*2D Line Plot*

2D line plots represent graphically the values of an array on the y-axis, at regular intervals on the x-axis (Fig. 15.9).

*Probability Plots (Quantile-Normal Plot/QN Plot, Quantile-Quantile Plot/QQ Plot)*

Probability plots are a graphical test for assessing if some data follows a particular distribution. They are most often used for testing the normality of a data set, as many statistical tests have the assumption that the exposure variables are approximately normally distributed. These plots are also used to examine residuals in models that rely on the assumption of normality of the residuals (ANOVA or regression analysis for example).

The interpretation of a QN plot is visual (Fig. 15.10): either the points fall randomly around the line (data set normally distributed) or they follow a curved pattern instead of following the line (non-normality). QN plots are also useful to identify skewness, kurtosis, fat tails, outliers, bimodality etc.



**Fig. 15.10** Example of QQ plot

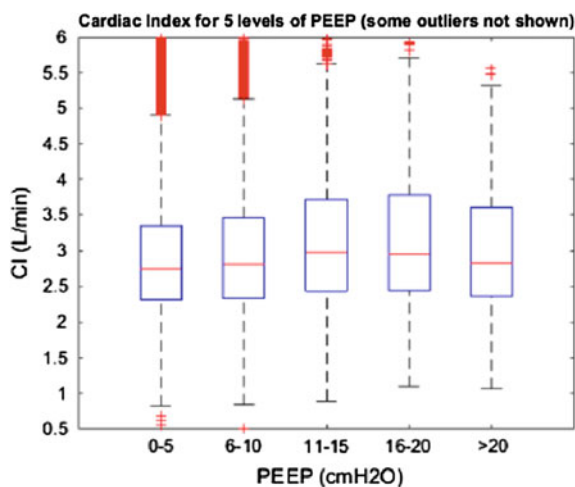
Besides the probability plots, there are many quantitative statistical tests (not graphical) for testing for normality, such as Pearson  $\chi^2$ , Shapiro-Wilk, and Kolmogorov-Smirnov.

Deviation of the observed distribution from normal makes many powerful statistical tools useless. Note that some data sets can be transformed to a more normal distribution, in particular with log-transformation and square-root transformations. If a data set is severely skewed, another option is to discretize its values into a finite set.

### ***Multivariate Graphical EDA***

#### *Side-by-Side Boxplots*

Representing several boxplots side by side allows easy comparison of the characteristics of several groups of data (example Fig. 15.11). An example of such boxplot is shown in the case study.



**Fig. 15.11** Side-by-side boxplot showing the cardiac index for five levels of Positive end-expiratory pressure (PEEP)

Scatterplots

Scatterplots are built using two continuous, ordinal or discrete quantitative variables (Fig. 15.12). Each data point's coordinate corresponds to a variable. They can be complexified to up to five dimensions using other variables by differentiating the data points' size, shape or color.

Scatterplots can also be used to represent high-dimensional data in 2 or 3D (Fig. 15.13), using T-distributed stochastic neighbor embedding (t-SNE) or principal component analysis (PCA). t-SNE and PCA are dimension reduction features used to reduce complex data set in two (t-SNE) or more (PCA) dimensions.

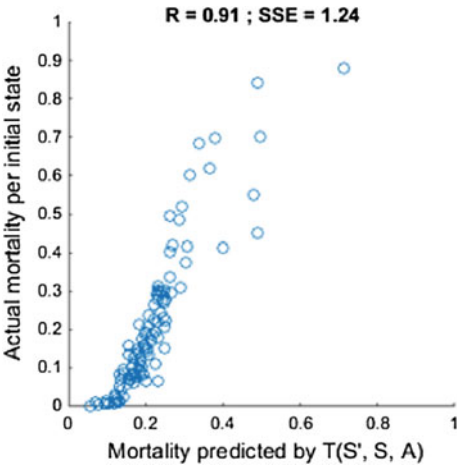


Fig. 15.12 Scatterplot showing an example of actual mortality per rate of predicted mortality

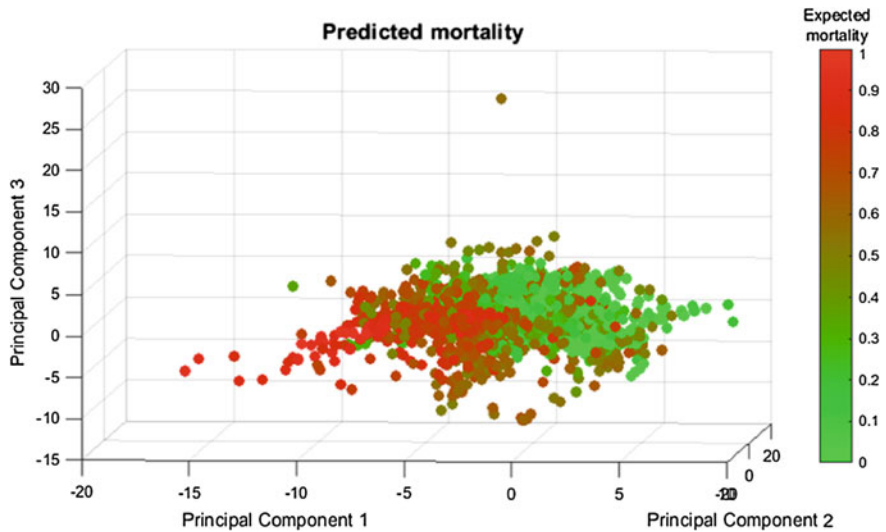


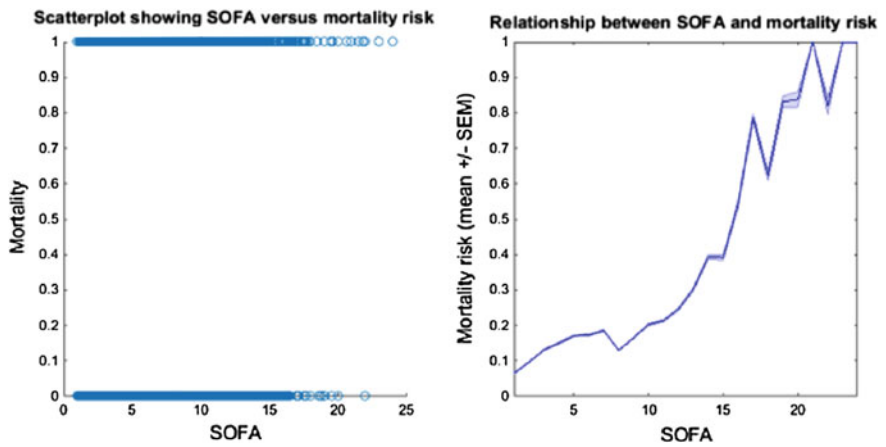
Fig. 15.13 3D representation of the first three dimension of a PCA



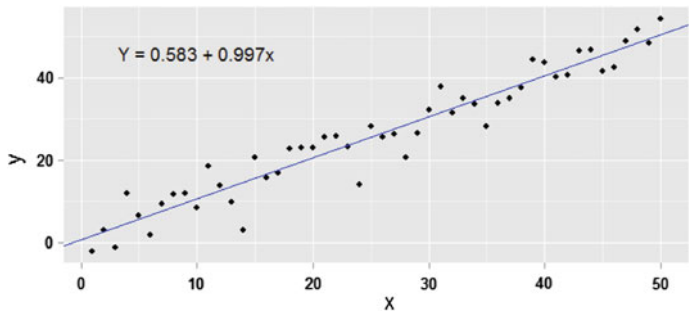
For binary variables (e.g. 28-day mortality vs. SOFA score), 2D scatterplots are not very helpful (Fig. 15.14, left). By dividing the data set in groups (in our example: one group per SOFA point), and plotting the average value of the outcome in each group, scatterplots become a very powerful tool, capable for example to identify a relationship between a variable and an outcome (Fig. 15.14, right).

*Curve Fitting*

Curve fitting is one way to quantify the relationship between two variables or the change in values over time (Fig. 15.15). The most common method for curve fitting relies on minimizing the sum of squared errors (SSE) between the data and the



**Fig. 15.14** Graphs of SOFA versus mortality risk



**Fig. 15.15** Example of linear regression

fitted function. Please refer to the “Linear Fit” function to create linear regression slopes in R.

### More Complicated Relationships

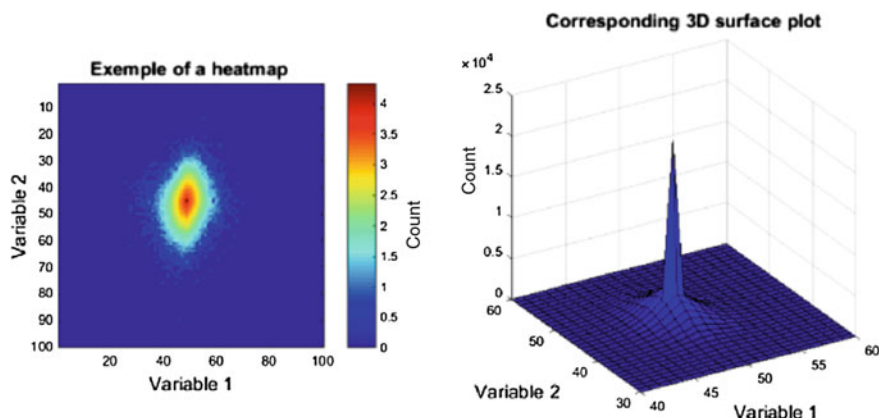
Many real life phenomena are not adequately explained by a straight-line relationship. An always increasing set of methods and algorithms exist to deal with that issue. Among the most common:

- Adding transformed explanatory variables, for example, adding  $x^2$  or  $x^3$  to the model.
- Using other algorithms to handle more complex relationships between variables (e.g., generalized additive models, spline regression, support vector machines, etc.).

### Heat Maps and 3D Surface Plots

Heat maps are simply a 2D grid built from a 2D array, whose color depends on the value of each cell. The data set must correspond to a 2D array whose cells contain the values of the outcome variable. This technique is useful when you want to represent the change of an outcome variable (e.g. length of stay) as a function of two other variables (e.g. age and SOFA score).

The color mapping can be customized (e.g. rainbow or grayscale). Interestingly, the Matlab function *imagesc* scales the data to the full colormap range. Their 3D equivalent is mesh plots or surface plots (Fig. 15.16).



**Fig. 15.16** Heat map (*left*) and surface plot (*right*)

## 15.3 Part 2—Case Study

This case study refers to the research that evaluated the effect of the placement of indwelling arterial catheters (IACs) in hemodynamically stable patients with respiratory failure in intensive care, from the MIMIC-II database.

For this case study, several aspects of EDA were used:

- The categorical data was first tabulated.
- Summary statistics were then generated to describe the variables of interest.
- Graphical EDA was used to generate histograms to visualize the data of interest.

### 15.3.1 Non-graphical EDA

#### *Tabulation*

To analyze, visualize and test for association or independence of categorical variables, they must first be tabulated. When generating tables, any missing data will be counted in a separate “NA” (“Not Available”) category. Please refer to the Chap. 13 “Missing Data” for approaches in managing this problem. There are several methods for creating frequency or contingency tables in R, such as for example, tabulating outcome variables for mortality, as demonstrated in the case study. Refer to the “Tabulate” function found in the GitHub repository for this book (see code appendix at the end of this Chapter) for details on how to compute frequencies of outcomes for different variables.

#### *Statistical Tests*

Multiple statistical tests are available in R and we refer the reader to the Chap. 16 “Data Analysis” for additional information on use of relevant tests in R. For examples of a simple Chi-square...” as “For examples of a simple Chi-squared test, please refer to the “Chi-squared” function found in the GitHub repository for this book (see code appendix at the end of this Chapter). In our example, the hypothesis of independence between expiration in ICU and IAC is accepted ( $p > 0.05$ ). On the contrary, the dependence link between day-28 mortality and IAC is rejected.

#### *Summary statistics*

Summary statistics as described above include, frequency, mean, median, mode, range, interquartile range, maximum and minimum values. An extract of summary statistics of patient demographics, vital signs, laboratory results and comorbidities, is shown in Table 6. Please refer to the function called “EDA Summary” in the

**Table 15.5** Comparison between the two study cohorts (subsample of variables only)

Variables	Entire Cohort (N = 1776)		
	Non-IAC	IAC	p-value
Size	984 (55.4 %)	792 (44.6 %)	NA
Age (year)	51 (35–72)	56 (40–73)	0.009
Gender (female)	344 (43.5 %)	406 (41.3 %)	0.4
Weight (kg)	76 (65–90)	78 (67–90)	0.08
SOFA score	5 (4–6)	6 (5–8)	<0.0001
<i>Co-morbidities</i>			
CHF	97 (12.5 %)	116 (11.8 %)	0.7
...	...	...	...
<i>Lab tests</i>			
WBC	10.6 (7.8–14.3)	11.8 (8.5–15.9)	<0.0001
Hemoglobin (g/dL)	13 (11.3–14.4)	12.6 (11–14.1)	0.003
...	...	...	...

GitHub repository for this book (see code appendix at the end of this Chapter) (Table 15.5).

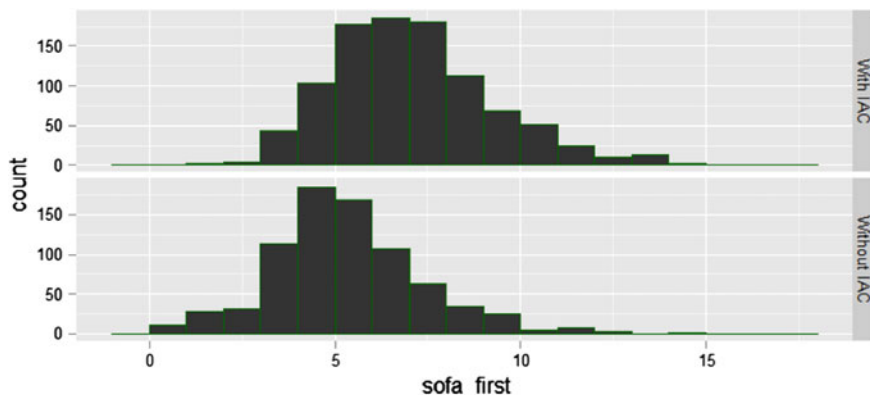
When separate cohorts are generated based on a common variable, in this case the presence of an indwelling arterial catheter, summary statistics are presented for each cohort.

It is important to identify any differences in subject baseline characteristics. The benefits of this are two-fold: first it is useful to identify potentially confounding variables that contribute to an outcome in addition to the predictor (exposure) variable. For example, if mortality is the outcome variable then differences in severity of illness between cohorts may wholly or partially account for any variance in mortality. Identifying these variables is important as it is possible to attempt to control for these using adjustment methods such as multivariable logistic regression. Secondly, it may allow the identification of variables that are associated with the predictor variable enriching our understanding of the phenomenon we are observing.

The analytical extension of identifying any differences using medians, means and data visualization is to test for statistically significant differences in any given subject characteristic using for example Wilcoxon-Rank sum test. Refer to Chap. 16 for further details in hypothesis testing.

### 15.3.2 Graphical EDA

Graphical representation of the dataset of interest is the principle feature of exploratory analysis.



**Fig. 15.17** histograms of SOFA scores by intra-arterial catheter status

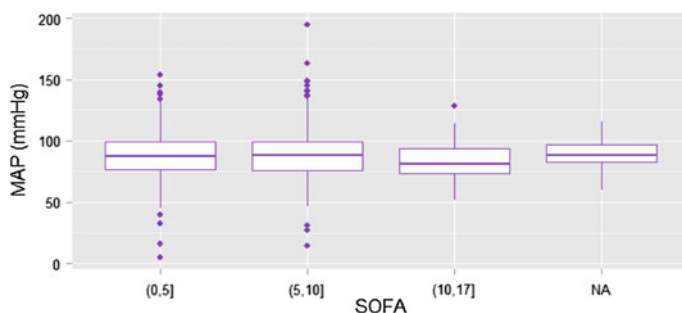
## Histograms

Histograms are considered the backbone of EDA for continuous data. They can be used to help the researcher understand continuous variables and provide key information such as their distribution. Outlined in *noise and outliers*, the histogram allows the researcher to visualize where the bulk of the data points are placed between the maximum and minimum values. Histograms can also allow a visual comparison of a variable between cohorts. For example, to compare severity of illness between patient cohorts, histograms of SOFA score can be plotted side by side (Fig. 15.17). An example of this is given in the code for this chapter using the “side-by-side histogram” function (see code appendix at the end of this Chapter).

## Boxplot and ANOVA

Outside of the scope of this case study, the user may be interested in analysis of variance. When performing EDA and effective way to visualize this is through the use of boxplot. For example, to explore differences in blood pressure based on severity of illness subjects could be categorized by severity of illness with blood pressure values at baseline plotted (Fig. 15.18). Please refer to the function called “Box Plot” in the GitHub repository for this book (see code appendix at the end of this Chapter).

The box plot shows a few outliers which may be interesting to explore individually, and that people with a high SOFA score (>10) tend to have a lower blood pressure than people with a lower SOFA score.



**Fig. 15.18** Side-by-side boxplot of MAP for different levels of severity at admission

## 15.4 Conclusion

In summary, EDA is an essential step in many types of research but is of particular use when analyzing electronic health care records. The tools described in this chapter should allow the researcher to better understand the features of a dataset and also to generate novel hypotheses.

### Take Home Messages

1. Always start by exploring a dataset with an open mind for discovery.
2. EDA allows to better apprehend the features and possible issues of a dataset.
3. EDA is a key step in generating research hypothesis.

**Open Access** This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, duplication, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, a link is provided to the Creative Commons license and any changes made are indicated.

The images or other third party material in this chapter are included in the work's Creative Commons license, unless indicated otherwise in the credit line; if such material is not included in the work's Creative Commons license and the respective action is not permitted by statutory regulation, users will need to obtain permission from the license holder to duplicate, adapt or reproduce the material.

## Code Appendix

The code used in this chapter is available in the GitHub repository for this book: <https://github.com/MIT-LCP/critical-data-book>. Further information on the code is available from this website.

## References

1. Natrella M (2010) NIST/SEMATECH e-Handbook of Statistical Methods. NIST/SEMATECH
2. Mosteller F, Tukey JW (1977) Data analysis and regression. Addison-Wesley Pub. Co., Boston
3. Tukey J (1977) Exploratory data analysis. Pearson, London
4. Seltman HJ (2012) Experimental design and analysis. Online <http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf>
5. Kaski, Samuel (1997) “Data exploration using self-organizing maps.” *Acta polytechnica scandinavica: Mathematics, computing and management in engineering series no. 82. 1997.*
6. Hill T, Lewicki P (2006) Statistics: methods and applications: a comprehensive reference for science, industry, and data mining. StatSoft, Inc., Tulsa
7. CRAN (2016) The Comprehensive R archive network—packages. Contributed Packages, 10 Jan 2016 [Online]. Available: <https://cran.r-project.org/web/packages/>. Accessed: 10 Jan 2016
8. Grubbs F (1969) Procedures for detecting outlying observations in samples. *Technometrics* 11(1)
9. Joanes DN, Gill CA (1998) Comparing measures of sample skewness and kurtosis. *The Statistician* 47:183–189.